DOCUMENT RESUME

ED 126 122 TM 005 377

AUTHOR	Swezey, Robert W.; Pearlstein, Richard B.
TITLE	Guidebook for Developing Criterion-Referenced Tests.
INSTITUTION	Applied Science Associates, Inc., Reston, Va.
SPONS AGENCY	Army Research Inst. for the Behavioral and Social Sciences, Arlington, Va.
REPORT NO	287-AR18(2)-IR-0974
PUB DATE	Aug 75
CONTRACT	DAHC-19-74-C-0018
NOTE	187p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
EDRS PRICE	MF-$0.83 HC-$10.03 Plus Postage.
DESCRIPTORS	Behavioral Objectives; Check Lists; Comparative Analysis; *Criterion Referenced Tests; Educational Objectives; *Guides; Item Analysis; Item Banks; Item Sampling; Norm Referenced Tests; Performance Tests; Scoring; Standards; *Test Construction; Test Reliability; Test Validity

ABSTRACT

This manual outlines the rationale for using the Criterion Referenced Test (CRT) approach and suggests specific guidelines for test developers to use in constructing test items. Methods for assessing the adequacy of a CRT are also provided. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original. *
****

ED126122

AD A014987

TM

GUIDEBOOK FOR DEVELOPING CRITERION-REFERENCED TESTS

Robert W. Swezey and Richard B. Pearlstein
Applied Science Associates, Inc.

NIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

U. S. Army
2
Research Institute for the Behavioral and Social Sciences

August 1975

Approved for public release; distribution unlimited.

TM005 377

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

W. C. MAUS
COL, GS
Commander

J. E. UHLANER
Technical Director

Research accomplished
under contract to the Department of the Army

Applied Science Associates, Inc.

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 1300 Wilson Boulevard, Arlington, Virginia 22209.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N/A	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) GUIDEBOOK FOR DEVELOPING CRITERION-REFERENCED TESTS		5. TYPE OF REPORT & PERIOD COVERED Manual 6. PERFORMING ORG. REPORT NUMBER 287-AR18(2)-1R-0974
7. AUTHOR(s) Robert W. Swezey and Richard B. Pearlstein		8. CONTRACT OR GRANT NUMBER(s) DAHC-19-74-C-0018
9. PERFORMING ORGANIZATION NAME AND ADDRESS Applied Science Associates, Inc. Reston International Center Reston, Virginia 22091		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 20164715A757
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral & Social Sciences, 1300 Wilson Blvd., Arlington, Virginia 22209		12. REPORT DATE August 1975 13. NUMBER OF PAGES 210
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Criterion Referenced Tests (CRT)
Test Development
CRT Construction

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This manual outlines the rationale for using the CRT approach and suggests specific guidelines for test developers to use in constructing test items. Methods for assessing the adequacy of a CRT are also provided.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Army Project Number
2Q164715A757

Unit Training Standards
and Evaluation

GUIDEBOOK FOR DEVELOPING CRITERION-REFERENCED TESTS

Robert W. Swezey

and

Richard B. Pearlstein
Applied Science Associates, Inc.

Angelo Mirabella, Work Unit Leader

Submitted by:
Frank J. Harris, Chief
Unit Training and Educational Technology Systems Technical Area

August 1975

Approved by:

Joseph Zeidner, Director
Organizations & Systems
Research Laboratory

J. E. Uhlaner, Technical Director
U.S. Army Research Institute for
the Behavioral and Social Sciences

Approved for public release; distribution unlimited.

FOREWORD

This publication is part of a larger program on criterion-referenced testing (CRT) being conducted by the Unit Training and Educational Technology Systems Technical Area of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The need for mastery-based performance testing is motivated by the need to differentiate students who can successfully demonstrate the required proficiency on a task from those students who cannot demonstrate the required proficiency. Progress in the application of CRT techniques has been impeded by the lack of easy-to-follow guidelines for test developers. A major goal of the program is to develop procedures for applying CRT theory and to evaluate the adequacy of the CRT approach in a variety of training situations. Related efforts in the Technical Area include: scoring procedures for performance-based training in tank gunnery (IDOC), experiments to compare the accuracy of several CRT models in fitting empirical data (METTEST), and the systematic development of training and testing objectives for tank gunnery (LIVEFIRE).

This publication outlines the rationale for using the CRT approach and suggests specific guidelines for test developers in constructing the test items. Methods for assessing the adequacy of a CRT are also provided.

ARI research in this area is conducted as an in-house effort augmented by contracts with organizations selected as having unique capabilities and facilities for research in a specific area. The present study was conducted by personnel of the Army Research Institute and Applied Sciences Associates, Inc., under Contract Number DAHC-19-74-C-0018, and is responsive to the requirements of RDTE Project 2Q164715A757, Training Systems Applications, FY 74.

J. E. Uhlaner

J. E. UHLANER,
Technical Director

GUIDEBOOK FOR DEVELOPING CRITERION-REFERENCED TESTS

CONTENTS

Page

CHAPTER 1--INTRODUCTION 1-1

HOW TO USE THIS MANUAL 1-1

PURPOSE 1-3

WHEN TO USE CRTs 1-5

CRT or NRT? 1-6

OTHER USES OF CRTs 1-8

Screening Devices 1-8
Diagnostic Aids 1-8
Evaluation of Instruction 1-9

OVERVIEW OF CRT CONSTRUCTION PROCESS 1-9

ESSENTIAL STEPS 1-11

CHAPTER 2--ASSESSING INPUTS TO THE CRT DEVELOPMENT PROCESS 2-1

LEVELS OF OBJECTIVES 2-1

THE THREE MAIN PARTS OF OBJECTIVES 2-3

Performance 2-4
Conditions 2-5
Standards 2-6
Separating Objectives Into Their Three Parts 2-8

ASSESSING THE ADEQUACY OF THE OBJECTIVES 2-9

Checking That Objectives are Unitary 2-10
Checking for Clarity of Main Intents 2-12
Ensuring That Performance Indicators Are Simple, Direct, and Part of the Trainees' Repertoire of Behavior 2-15
Checking That Performances, Conditions, and Standards are Specified in Precise, Operational Terms 2-17
Summary 2-20

CONTENTS (continued)

Page

CHAPTER 3--DEVELOPING A TEST PLAN 3-1

EXAMINING PRACTICAL CONSTRAINTS 3-1

Time 3-2
Manpower 3-2
Costs 3-3
Facilities/Equipment 3-3
Degree of Realism 3-4
Potential Sources of Data 3-5
Assessing Practical Constraints 3-5
Selecting Among Objectives 3-6
Modifying Objectives in Light of Practical Constraints 3-7
Submit Modified Objectives 3-10

PLANNING ITEM FORMAT AND LEVEL OF FIDELITY 3-12

Types of Items for Written Tests 3-14
Items for Performance Tests: Process and Product Measures 3-16
Types of Items for Performance Tests: Process Rating 3-19
Types of Items for Performance Tests: Product Rating 3-23
Example of Determining Item Format and Test Fidelity 3-24

ITEM SAMPLING AND SAMPLING AMONG CONDITIONS 3-25

Should Performances be Tested Under Single or Under Multiple Conditions 3-27

DETERMINING HOW MANY ITEMS TO INCLUDE IN YOUR TEST, AND DOCUMENTING YOUR TEST PLAN 3-29

The Test Plan Worksheet 3-31

CHAPTER 4--CONSTRUCTING THE ITEM POOL 4-1

CREATE ITEMS BASED ON TEST PLAN SPECIFICATIONS 4-1

DEVELOP AND DOCUMENT INSTRUCTIONS FOR ITEM USE 4-6

ASSESSING ADEQUACY OF ITEMS 4-8

Do Items Match Objectives? 4-8
Other Checks on Item Adequacy 4-9

DEVELOP GENERAL TEST INSTRUCTIONS 4-10

CONTENTS (continued)

Page

CHAPTER 5--SELECTING FINAL TEST ITEMS 5-1

TRYING OUT THE ITEM POOL 5-1

Selecting A Sample 5-1
Sample Size 5-2
Determination of Test Tryout Samples: Illustrative Problem 5-5
Solution 5-5
Conducting a Tryout 5-6
Conducting An Item Analysis On the Tryout Results 5-7
Calculating ϕ 5-7
Using ϕ 5-12
Summary of Using ϕ In Item Analysis 5-13
Other Points About Item Analysis 5-14
Item Analysis by Inspection 5-14
Cautions on Use of Item Analysis Techniques 5-15

REVIEWING REMAINING TEST ITEMS 5-16

Feedback From Individuals in the Tryout Sample 5-16
Peer Review 5-18
Formal Review by Test Evaluation Units 5-19
Formal Review by Subject Matter Experts 5-19

REDUCING THE ITEM POOL 5-19

What To Do If You Eliminate Too Few Or Too Many Items 5-21

CHAPTER 6--ADMINISTERING AND SCORING CRTs 6-1

CONTROLLING THE TEST SITUATION 6-1

Controlling Environmental Variables 6-1
Controlling Personal Variables 6-2
Instructions and Tester Variables 6-2

SCORING PROCEDURES 6-5

Assist vs. Non-Interference Scoring 6-6
"Go - No-Go" Scoring 6-7
Fixed Point Scoring 6-7
Mixed Scoring Techniques 6-8
Rating Scales 6-9
Establishing Cut-Off Scores 6-11
False Positives and False Negatives 6-12

CONTENTS (continued)

Page

REPORTING AND RECORDING TEST RESULTS 6-14

SPECIAL PROBLEMS 6-14

CHAPTER 7--ASSESSING RELIABILITY AND VALIDITY 7-1

ASSESSING RELIABILITY 7-2

Computing ϕ as an Estimate of Reliability 7-3

ASSESSING VALIDITY 7-6

Determining Content Validity 7-7
Determining Concurrent Validity 7-9
Determining Predictive Validity 7-11

APPENDIX A--CHECKLIST FOR CONSTRUCTING CRTs A-1

APPENDIX B--CHECKLIST FOR EVALUATING CRTs B-1

APPENDIX C--GLOSSARY C-1

APPENDIX D--SQUARE ROOT TABLES D-1

APPENDIX E--REVIEW QUESTIONS AND ANSWERS E-1

FIGURES

			Page
Figure	1-1	Comparison of CR Testing with NR Testing	1-5
	1-2	Sample Test Results	1-7
	2-1	Some Synonyms for the Parts of an Objective	2-3
	2-2	Six Types of Standards	2-7
	2-3	Sequence of Operations for Assessing the Adequacy of an Objective	2-21
	2-4	Examples of Verbs Often Used to Specify Performance in Objectives	2-18
	2-5	Examples of Statements of Conditions and Standards	2-19
	3-1	Sequence of Operations for Developing a Test Plan	3-35
	3-2	Guideline for Selecting Among Objectives in CRT Development	3-7
	3-3	Tabular Form for Summarizing Conditions and Standards that Require Change in an Objective and How to Change Them	3-11
	3-4	Fidelity Levels and Types of Measurement	3-13
	3-5	Some Common Differences Between Performance Test Items and Written Test Items	3-16
	3-6	Sample Numerical Scale for Rating Public Speaking Ability	3-19
	3-7	Sample Behaviorally-Anchored Rating Scale	3-20
	3-8	Sample Numerical Scale for Rating Driving a Truck	3-20
	3-9	Sample Behaviorally-Anchored Rating Scale	3-24
	3-10	Multiple Testing Conditions	3-28
	3-11	Test Plan Worksheet	3-33
	3-12	Sample Test Plan Worksheet	3-34

FIGURES (continued)

Page

Figure 4-1 Sequence of Operations for Constructing the Item Pool 4-13

4-2 Sample Multiple Choice Test 4-3

4-3 Sample Illustrated Multiple-Choice Test 4-4

4-4 Sample Simulated Performance Test 4-5

5-1 Sequence of Operations for Selecting Final Test Items 5-25

5-2 Guidelines for Choosing Sample Size 5-4

5-3 Results of Item Tryout 5-8

5-4 Organization of Tryout Results for Computing φ for Item 4 5-9

5-5 Item/Test Matrices Filled in for the Tryout Results Shown in Figure 5-3 5-10

5-6 Formula for φ 5-11

5-7 Range of Values of φ 5-12

5-8 Values of φ for Items in Tryout Sample 5-12

5-9 Worksheet for Recording Feedback from Tryout 5-17

5-10 Item Pool Review Summary Sheet 5-20

5-11 Item Pool Review Summary Sheet with Sample Entries for a 10-Item Pool 5-21

5-12 Alternate Test Forms Possible for a Four-Item Test Made from Six Items 5-22

6-1 Sequence of Operations for Administering and Scoring CRTs 6-17

6-2 Typical Test Instructions 6-3

6-3 Some Typical Testing Steps 6-4

6-4 Three Components of the Test Situation 6-5

6-5 Comparison of Ratings on a 6-Item Test 6-10

6-6 Types of CRT Scoring 6-11

6-7 False Positives and False Negatives 6-13

FIGURES (continued)

		Page
Figure 7-1	Sequence of Operations Involved in Assessing Reliability and Validity	7-15
7-2	Matrix Used for Computing ϕ in Test-Retest Reliability Estimates	7-4
7-3	Matrices for Test-Retest Reliability Estimates with Sample Data for Two Different Tests	7-5
7-4	Three Types of Validity	7-6
7-5	A One-Item CRT and Its Objective	7-7
7-6	Matrix for Concurrent Validation with Sample Data	7-10

CHAPTER 1

INTRODUCTION

HOW TO USE THIS MANUAL

This manual is intended for use by persons involved in testing. You will find this manual useful if your work involves any phase of test construction, test administration, or use of test results. Whether you are involved with just a small segment of test construction and use--such as writing a few test items or helping administer performance tests at a field-station--or whether you supervise an entire test construction or test administration process, you will find helpful guidance in this manual.

This manual is a carefully researched presentation of what is known about Criterion-Referenced (CR) testing, written in a "how-to-do-it" fashion. Examples used in this manual to illustrate points are drawn from the experience of Army test personnel working in a variety of Army situations. Although test construction and use requirements differ in various Army facilities, this manual has been tailored to be as useful to you as possible, no matter what particular processes are used to develop and administer tests at your location. Consequently, while this manual does present an overall procedure for developing and using Criterion-Referenced Tests (CRTs), it is not essential that you follow all steps just as they are presented here. Rather, you can use this manual for guidance in performing particular steps, without violating the overall way in which you develop tests. Of course, if you follow the overall process presented in this manual, you can be more certain that you will develop tests that will measure what you want them to measure.

While there are certain technical questions involving CRT construction on which testing experts fail to agree, there is basic agreement on many major elements. So, if you are presently involved with test development and use, you will find in this manual guidelines that can help you in performing your particular testing tasks, steer you around problems, and help ensure that your tests work as well as possible.

The emphasis in this manual is on test development. If you are involved only in the administration of tests, you might want to read just Chapter 6 which covers administering and scoring of tests. If you are involved in only a small segment of an overall test construction effort, or if you have a problem with a specific aspect of test development, you may just want to consult the relevant section of this manual. Refer to the table of contents to find the appropriate reading to aid you.

Although this manual tries to avoid the use of technical testing terminology, you may find some terms that are unclear to you. You can use the glossary in Appendix C of this manual for help in such cases.

After you are familiar with the test construction processes contained in this manual, you may wish to use a checklist to guide you in your test development activities. The Checklist for Constructing CRTs contained in Appendix A of this manual will help you ensure that all steps in the test construction process are covered adequately.

If you are concerned with assessing CRTs that have already been built, you may want to use the Checklist for Evaluating CRTs to guide you in your evaluation. You may also want to use this checklist as a guide for reviewing tests you build prior to formal tryout. This checklist appears in Appendix B.

The following features help make this manual easy to use:

- Review questions and answers for each chapter (in Appendix E) will help you to supplement your depth of understanding for that chapter.
- Pages are numbered within chapters.
- Chapters have flowcharts when necessary to show the sequence of operations required for completing CRT development tasks. The flowcharts fold out so you can refer to them as you read the text. By using these flowcharts, you can see just where you are in the CRT development process.
- Major points are surrounded by boxes, and other points are identified by bullets (•) to make them stand out.
- Examples are highlighted for easy reference.

PURPOSE

The purpose of this manual is to provide guidance on the construction and use of Criterion-Referenced Tests (CRTs). CRTs are relative newcomers to the field of testing. Because of their advantages, CRTs are receiving ever-increasing application. You may already be involved with CRTs, without realizing it, since there is still some disagreement in terminology. For example, many so called "performance tests" are actually CRTs. To clear up any confusion, let us define Criterion-Referenced testing (CR testing). A CRT measures what an individual can do or knows, compared to what he must be able to do or must know in order to successfully perform a task. Basically, what this means is that an individual's performance is compared to (referenced to) some external criteria, or performance standards. These standards are derived from an analysis of what is required to do a particular task successfully.

The traditional approach to testing is called Norm-Referenced testing (NR testing). In NR testing, an individual's performance is compared to the performance of other individuals. For example, any time your test is scored "on a curve," your performance is being compared to that of others. Suppose an individual takes a NRT on his ability to repair a 2-1/2 ton truck transmission and scores at the 90th percentile. At best, all this tells you is that the individual can repair such a transmission better than 90 out of 100 other individuals who take the test. It does not tell you that the individual can repair this transmission to specific test standards--that he can fix it so that it will work and hold up for a reasonable period of time under normal operating conditions. A CRT on the same subject would tell you whether or not the individual could repair the transmission to the appropriate standards. Scores from this CRT might be recorded in terms of "go" or "no-go." All individuals who received a "go" (or a "pass") on the CRT, would be able to repair the 2-1/2 ton truck transmission to the test standards. You would not necessarily know whether one individual who got a "go" did better work than another who also got a "go," but you would know that both had enough knowledge and skill to repair such transmissions.

In many cases, you can't tell a CRT from a NRT just by looking at the test: the items on both tests might look the same. Both CRTs and NRTs may have multiple-choice items or fill-in-the-blank items. They both may use simulated performance measures such as:

- Tie the tourniquet on the dummy's leg
- Demonstrate proper bayonet procedures using the rubber mock-up M-16 and bayonet

or hands-on performance measures such as:

- Disassemble this weapon
- Connect the calling party to the called party using standard field switchboard

Both CRTs and NRTs may have knowledge-type items such as:

- At what temperature should a layer cake be baked?
- What symptoms indicate that an atropine injection should be administered?

or skill-type items such as:

- Compute the elevation required for firing a howitzer round from point X to specified grid coordinates
- Replace the faulty component on this radio chassis

Both types of tests may have paper-and-pencil performance items. For example:

- Plot the quickest route from point A to point B on the topographic map supplied.

or actual performance items. For example:

- You are dropped at point A in the test range. Using the map and magnetic compass provided, [illegible] to point B within two hours.

So, looking at a test will not necessarily tell you whether or not it is a CRT. To determine if a test is a CRT, you need to find out how it was developed, what it is used for, and how the score is interpreted. A test is criterion-referenced if:

- The test items are based upon training objectives which, in turn, were developed from performance objectives external to training. That is, the development of the test can be directly traced to a consideration of the tasks which the trainee will eventually perform on the job.
- The test is primarily used for measuring mastery. That is, the test is designed to determine whether or not the individual has mastered particular tasks. CRTs may also be used to assess instructional programs; that is, they may help determine whether or not programs do train individuals to achieve mastery.
- Scoring of the test is based upon absolute standards such as job competence rather than upon relative standards such as class standing.

If a test meets the above three criteria, it is criterion-referenced.

Figure 1-1 presents a summary of CR testing as compared to NR testing. As you can see from this table, only by using CR testing can you know whether an individual is prepared to do a job. NRTs may be able to tell you which individuals are more prepared than others, but not which are ready to do the job.

CR TESTING	NR TESTING
Requires a careful analysis of skills and knowledges needed for performing tasks on which individuals are to be tested--Task analytic data provide the basis for the construction of items	May be based on course content taught or instructor's assumptions of what individuals need to know--Task analytic data are not necessarily considered
Test results indicate whether or not an individual can perform a task to acceptable standards	Test results indicate how well an individual does (or how much an individual knows) as compared to others who have taken the test
Test results are most useful for making absolute decisions, such as whether or not a person is ready to perform a particular job task	Test results are most useful for making relative decisions such as who knows more, who works more quickly, or class rank

Figure 1-1. Comparison of CR Testing with NR Testing

WHEN TO USE CRTs

You can develop and use CRTs for a variety of purposes. The foremost use of CRTs is to answer the question "How well can the individual perform compared to how well he needs to perform to accomplish a task?" In other words, you should use a CRT whenever you need to find out if an individual knows and can do what is required in order to perform the tasks for which he is being trained.

Remember though, you'll have to be able to meet two other criteria, aside from the answer to the above question, before you can build a CRT.

- First, you'll have to be able to base your test items on training objectives which were developed from performance objectives external to training. So, if you can't point to external performance objectives (what the individual should be able to do on the job after training), you can't develop a CRT that will be a useful measure of job performance.

- Second, you'll have to be able to score the test on an absolute basis. If the test won't be scoreable in this way--that is, if you can't specify the minimum acceptable standards for adequate performance--then you won't be able to build a CRT.

A properly constructed CRT will allow you to classify the people who take it into two groups:

- Masters--those who you are reasonably sure can do what they are trained for,

and

- Non-masters--those who you are reasonably sure cannot adequately do what they are trained for.

A CRT, then, lets you find out whether or not an individual has mastered a task or skill.

If you are interested in finding out who does best, who does average, and who does worst, you should not use a CRT. In fact, whenever you want to answer the question "How well does an individual do compared to others?", you should use a NRT instead of a CRT. NRTs are designed to produce large differences in the scores of people taking them, so they can be used for helping you find out who does best, second best, third best, etc. CRTs, though, usually don't produce large score differences--all masters may get just about the same score--so they are not good for helping you put people in the order of how well they do compared to one another.

CRT or NRT?

Suppose you wanted to test a class at the end of training and name the two top scorers as honor graduates.

- Question--Would you want to give the class a CRT or a NRT?

If you give a CRT, you may find that most of the class are masters--most of the class can do what you have trained them for. But 10 people in your class may get the top score, so which two do you name as honor graduates? On the other hand, if you give them a NRT, you will probably find two individuals who clearly score higher than the rest of the people in the class. But, with a NRT all you know is that these individuals do best compared to the other people who took the test. You don't necessarily know whether or not these two have mastered the tasks. Just the same, you would have a clear basis for naming the two individuals who scored highest on the NRT as honor graduates. So, if you must name honor graduates (or select a few people for promotion or other special honors), you would be better off using a NRT. But if you want to find out who in your class has mastered the training, you had better use a CRT.

Now, suppose you receive a directive indicating that approximately five percent of your class are to be identified as honor graduates. You give the class a CRT which has a cut-off point at the score of 70. Anyone who scores 70 or above on the test has met the minimum acceptable standards on the tasks you've trained them to perform. Eighty is the top score possible--it represents perfect performance on the tasks tested.

There are 100 people in your class and they received the following scores:

Score	Number of people in class who get this score
80	20
78	40
77	10
76	10
75	5
74	5
72	5
71	5
	100

Figure 1-2. Sample Test Results

Now what do you do? Not only has everyone in your class passed the test, but 20 percent of the class have achieved perfect scores. Which people would you designate as honor graduates? You would have to find some way other than CRT scores to identify five percent of your class as honor graduates.

- So--if you need to use a CRT, you should not choose among class members on the basis of CRT results. All you can really say is who can do what they're supposed to, and who can't.

OTHER USES OF CRTS

Screening Devices

Another use of CRTs is as a screening device. If an individual needs to possess certain entry behaviors before he starts an advanced course, for example, you might want to give him a CRT before permitting him to start the course. In this case, the CRT would be based on objectives for tasks that the individual should be able to perform before beginning the course. A learner's permit test, for example, is often used as a screening device in automobile driver licensing: If an individual passes this test it means that he has the entry level knowledge--knowledge of state traffic laws--and can be considered ready to begin hands-on driver training.

You can also use a CRT as a screening device to see if the individual already knows how to perform some of the tasks. In some cases, an individual may be able to do a job without taking a training course because he has had appropriate past experience, or was trained for something similar. For cases like this, you might want to test this individual at the beginning of the course (or block of instruction; or sub course, or specialty area) with the same CRT you would give to the rest of the class at the end of the course (or block of instruction, etc.). If the individual achieves a mastery-level score on the test, then you won't have to waste resources or time by putting him through something that he can already do.

Diagnostic Aids

CRTs may also be useful as diagnostic aids. You can build a CRT so that it shows just what objectives an individual is weak on (has not yet mastered) and even what particular steps of a certain procedure he is unable to perform. A diagnostic CRT on drill and ceremonies, for example, may show that an individual cannot correctly execute "parade rest." By examining that individual's test score sheet, you might find that he failed parade rest because he did not hold his head and eyes at the position of attention. Remediation for this person becomes simple: You don't have to teach him all the steps of parade rest--his feet, arms and hands are in the correct positions--you only have to teach him to hold his head and eyes at attention. Of course, this is an overly simple example, but the principle holds true for much more complicated tasks, such as flying a helicopter.

Evaluation of Instruction

A final, major use of CRTs is to answer the question "Has my instructional program taught what it is supposed to teach?" That is, you can use a CRT to evaluate how good an instructional program is. If you have designed an instructional program to train people in specific tasks, you can use a CRT to find out how good the program is, as follows:

- First, you find an appropriate group of people who cannot do the tasks--the CRT should show that they are non-masters on those tasks.
- Then these people go through the instructional program.
- Finally, you test them with the CRT again.

If the instructional program is good, most should score as masters on the CRT after they've had the program.

OVERVIEW OF CRT CONSTRUCTION PROCESS

There is no single correct way to construct a CRT. The construction process outlined in this section is designed to help you construct and use CRTs that will be suitable for their intended applications. Following this process will help you cover all points necessary for an adequate test. However, your own imagination and ingenuity will be required to create workable tests. The process presented in this manual is designed to be applicable to diverse types of testing needs and situations, regardless of subject matter. Thus, you will need adequate knowledge of the subject matter or access to subject matter experts.

Remember, the CRT construction process presented here, is only one way of constructing and using CRTs. There may be other useful approaches which you have been following. Consequently, regard the information presented within the steps of this process as guidelines to aid you, not as absolute doctrine. If the process conflicts with your procedures, use only those guidelines which help you. If, on the other hand, you are starting from scratch in the test development process, you will find the CRT construction procedures presented here to be a simple and efficient method for constructing CRTs that will do the job. Here is a brief outline of the major steps for constructing, using and evaluating CRTs. They will be described in greater detail in Chapters 2 through 7.

1. Assessing Inputs to the CRT Development Process. In this step you assess the adequacy of the objectives that you will use in developing CRTs. Inadequate objectives must be revised or discarded. In assessing the adequacy of objectives, you will carefully consider their three main parts:

   - Performances--what the objective requires people to know and do.

   - Conditions--the situations under which people's performance will be evaluated.

   - Standards--the level of performance which indicates satisfactory achievement of the objective.

2. Developing a Test Plan. Before writing test items, you should plan the test. In this step you develop a test plan by considering a number of factors including:

   - Practical constraints--do factors such as time and manpower availability, costs, etc. affect the way the test must be built?

   - Item format--are the objectives best tested by written items, performance items, measures of how a performance is done, measures of products resulting from performance, etc.? How realistic should the test items be?

   - Number of items--how many items should be made for each objective? What kinds of conditions should the items include?

3. Constructing the Item Pool. In this step you create the items called for by your test plan. Wherever your test plan calls for one item, you create two. In this way you will create an item pool from which the best items can be selected by tryout and review procedures. After you have prepared all the items for your item pool, you assess the adequacy of each item considering such factors as:

   - Does it match the objective for which it was created?

   - Is it clear and unambiguous?

- Is it reasonably easy to administer?

- Is it at the appropriate level of realism as specified in the test plan?

In this step you also prepare instructions which tell how each item is to be administered. In addition, general instructions for the test as a whole must also be developed.

4. Selecting Final Test Items. In this step you try out the item pool and obtain reviews of the test items. Poor and redundant test items are revised or discarded, as necessary. You may also have to create and try out new items, if the first tryout and reviews eliminate items which leave gaps in the test plan.

5. Administering and Scoring the Test. In this step you create the scoring standards and administrative procedures for the test. You develop and document standardized conditions for test administration so the test can be administered and scored by others using your documentation. You also develop cut-off points for your test which tell what a passing score on the test (or on each of the objectives) is.

6. Measuring Reliability and Validity. In this step you evaluate the reliability of your test--that is, you find out if the test measures the same thing each time it is given. You also evaluate the validity of your test--that is, you determine whether or not it is actually measuring what it is supposed to measure. If your test has low reliability or validity you must consider ways of improving the test.

ESSENTIAL STEPS

Whether or not you use the CRT construction process step-for-step as described in the manual, you should be sure that the following essential points are covered in developing and using tests:

- Test items should be developed to reflect the attainment of objectives, which in turn are developed from independent analyses of the tasks. Test items should measure the performance specified in the objectives, under the appropriate conditions, to the specified standards.

- You should make sure that your test items meet the practical constraints of the training and testing situations, and that you try out your test items. Trying out items is the only certain way of finding out which items work best.

- You should review the results of the tryout and evaluate the items with peers, test evaluation units and subject-matter experts.

- You should provide appropriate administration and scoring procedures to be used with your CRT to ensure that the CRTs will be administered and used in a uniform and appropriate way.

CHAPTER 2

ASSESSING INPUTS TO THE CRT DEVELOPMENT PROCESS

The inputs to the CRT development process are called objectives. CRTs are developed from objectives that tell what people must do to successfully complete training or perform certain tasks. In this chapter we will first discuss different levels of objectives. Next, we will examine the three main parts which all objectives should include. Then we will see how to assess the adequacy of objectives. If objectives are inadequate, any test developed from them will be inadequate.

LEVELS OF OBJECTIVES

Objectives, and the CRTs which are developed from them, can be written at several different levels of detail. It's important to grasp what these levels are because they influence how tests are prepared and used. Understanding these levels can also help you judge the adequacy of the objectives from which tests must be derived.

Three basic levels can be identified:

- Level 1 refers to objectives which are prepared on the basis of doctrine and/or experience about actual, meaningful units of work activity which occur in operational environments. A number of different labels have been applied to such objectives including:

    - Job Performance Requirements (JPRs)
    - Performance Objectives (POs)
    - Performance Measures (PMs)
    - Job Objectives (JOs)
    - Task Objectives

  The exact labels are not important. What is important is knowing that Level 1 objectives refer to meaningful units of work activity performed under operational conditions, and according to operational standards. That is, Level 1 objectives tell what must be done on the job. The job-task analyst is principally responsible for such objectives.

- Level 2 objectives are essentially Level 1 objectives which have been modified by the training system or by the training program designer so that they match training resources and safety requirements. Level 2 still refers to meaningful units of work activity. Objectives in this category have been labeled:

    - Training Objectives
    - Instructional Objectives
    - Instructional Goals
    - Learning Objectives
    - Terminal Training Objectives

  This level describes work activities which can stand by themselves and still be meaningful. For example, operating a multimeter would be a Level 2 objective, if the intention were to train assembly line workers to perform quality checks to make sure that multimeters are operating properly before they are packaged for shipment. However, operating a multimeter is not necessarily a meaningful activity, apart from troubleshooting a malfunctioning electronic circuit. Operation of a multimeter in that case would be defined as a Level 3 objective, which will be described later.

  The point is: Level 2 objectives tell what a person must be able to do at the end of training, not necessarily in an operational environment (on the job). While the training program designer is principally responsible for these objectives, test developers have important contributions to make along with job task analysts and unit commanders. Testing at this level is designed to screen out individuals who have not mastered the objective(s) of a particular stage of training.

- Level 3 objectives refer to activities (component skills and knowledges) which are not directly useful by themselves. They are generated in an attempt to make training efficient and manageable. Labels used at this level include:

    - Enabling Objectives
    - Knowledges
    - Skills
    - Intermediate Objectives
    - Learning Elements (mental, physical, information, and attitude elements)

Level 3 objectives tell what a person must know and do as a prerequisite for doing Level 2 objectives. Testing at this level primarily serves a training and diagnostic purpose and is usually built into the training in the form of self quizzes.

THE THREE MAIN PARTS OF OBJECTIVES

Before constructing a CRT, it is necessary to take a close look at the objective(s) on which the CRT is to be based. You must thoroughly check each part of the objective. A properly written objective, regardless of level, should consist of the following three parts:

- Performance (Task)
- Conditions
- Standards

You are probably already familiar with these parts of an objective, but you may know them by other names. Figure 2-1 shows some of the other labels by which the main parts of objectives are identified.

Performance	Conditions	Standards
• Task	• Job conditions	• Training standard
• Action	• Environment*	• Criterion (plural= criteria)
• Skills, knowledges, and attitudes	• Tools and equipment*	• Job standards
• Subtask	• Working conditions*	• Pass/fail standards
• Objective (sometimes used as a label for performance only)	• Job aids*	• Go - no-go standards
	• Materials required*	
	• Notes*	

*These are specified kinds of conditions, all of which go to make up conditions as a whole.

Figure 2-1. Some Synonyms for the Parts of an Objective

Let us consider each of these main parts separately. After this we will look at examples of how to divide objectives into their three parts.

Performance

Every objective should state precisely what the individual must do. The statement of performance must be clear enough for that performance to be trained and tested. Examples of performances stated in objectives are:

- Climb the telephone pole
- Disassemble an M-16 rifle
- State the conditions for which a tourniquet should be applied
- Camouflage the helmet
- Add two five-digit numbers

Note that every statement of performance includes an action verb. This verb usually is the key to the performance. It tells what must be done. For example, in the statement of performance "State the conditions for which a tourniquet should be applied," the action verb is "state." You can test the student's ability to state these conditions. Suppose that statement of performance had read "Appreciate the conditions for which a tourniquet should be applied." Would you know what to test? How would you know when a student "appreciates" the conditions?

Sometimes, though, the action verb is not the key to the performance to be trained and tested. It may be only the indicator of the performance. Any time that you can't point to the performance itself, the action verb should specify the appropriate indicator of that performance. For example, consider the statement of performance "Add two five-digit numbers." It is clear that the performance called for is "adding." But how do you know when someone successfully adds two numbers? Obviously, an indicator must be supplied, since you can't observe the act of adding. So you would attach an indicator to the statement of performance; i.e., "Add two five-digit numbers and write the answer in the space below." Note that although "write. . ." is the observable action, the main intent of the performance is adding, not writing. If the statement of performance calls for an action (has a main intent) that is not directly observable, an appropriate indicator must be added. We will discuss main intents and indicators further in the next section, "Assessing the Adequacy of Objectives."

Conditions

Every objective should include a statement of the conditions under which the performance must be demonstrated. Such statements should indicate:

- What the student has to work with--what he is allowed to use (tools, reference materials, etc.)
- The environmental circumstances under which the performance must be demonstrated (nighttime conditions, classroom conditions, etc.)
- What the student must work on--his starting points (the "givens"--e.g., given a Mark II chassis. . .)
- Any limitations, special instructions, etc.

It is very important for an objective to specify all conditions which may affect performance. Without statements of these conditions, you can't be sure of just what to train or to test. Suppose, for example, that an objective stated "Be able to disassemble and reassemble an M-60 machine gun." You, the foot soldier, read the objective, receive training, and are ready to be tested. Your drill sergeant takes you into a windowless room, closes the door, hands you the machine gun, turns off the lights and says "Okay, disassemble and reassemble this weapon."

You say, "But Sergeant, the objective didn't say anything about doing it in the dark." He answers, "This is a combat weapon and you might have to use it anytime--night or day. I won't always be around to turn the lights on for you."

So, if conditions aren't specified, the student won't know exactly what he needs to learn to do. And, as a test developer, you won't know just what it is you should test. If you read the preceding objective, what conditions would you test under? Day? Night? Classroom? Rain? You would have to make an educated guess because you really wouldn't know.

Often performance must be demonstrated under multiple conditions. These must be specified. For example, if a student must learn to navigate through many different types of terrain, the objective should state each of the terrain conditions through which the student will have to find his way. Sometimes performance must be demonstrated under any of several conditions. In such cases, the statement of conditions in the objective should make clear that the performance need be demonstrated under only one of the conditions. For example, an objective requiring a trainee to determine the coordinates of a grid on a map may state "The trainee may do this indoors or outdoors."

The following statements represent some example conditions (statements of conditions are underlined):

- Given the volume of a sphere and the appropriate formula, compute the diameter of the sphere.
- Cross a standard obstacle course, in the rain.
- List the major components of a jeep clutch and their part numbers, using the reference manual provided.
- Replace the transistor on this circuit board without causing heat damage to the adjacent crystal diode.

Standards

Each objective should specify the standard (criterion) by which performance is evaluated.

In other words, every objective should indicate how well or how quickly (or both) a performance must be done. As is the case for statements of performance and conditions, standards, too, must be clearly stated in the objective or you won't know how to train or test. For example, suppose an objective only stated "Be able to type reasonably accurately using an electric typewriter under standard office conditions." Lacking standards for speed and accuracy, how fast would you train people to type in order to satisfy the objective? How fast would they have to type to pass a CRT? Obviously, the objective is lacking a clear statement of standards ("reasonably accurately" doesn't really tell you anything). A complete objective might read "Using an electric typewriter in standard office conditions, be able to type 50 words per minute corrected for accuracy (one word per minute subtracted for each mistake)." Working from such an objective you would know what standards to shoot for in training and the level of performance a person has to demonstrate on a test.

There are six specific types of standards that can be stated in objectives to indicate how well (quality) or how quickly (time) a performance must be done or a product completed. Figure 2-2 describes these types of standards. An objective should specify at least one of the six types of standards in order to be complete. Often an objective will combine several types of standards; for example, one of quality and one of time specifications.

Standard Refers to:	Type of Standard	Example (Statement of Standard Underlined)
Quality	Standard operating procedure--performance must match a specified SOP. This standard specifies that a performance be complete--all parts of performance done in sequence.	"Given a map with forward observers and enemy troop positions marked, the trainee must issue a "call-for-fire" using the proper sequence as specified in the Artillery Man's Handbook."
Quality	Zero error--performance must be completed to 100% accuracy (or product must be made exactly right).	"The trainee will set the quadrant on a M-102 Howitzer quadrant sight to a specified mil. He must set it at the exact mil (for example, 345) he is told"--If the trainee is off by one mil, he does not meet the standard.
Quality	Minimum acceptable level--performance must meet a specified minimum acceptable level (or product must meet specified tolerances).	"Using a standard oral thermometer, take a patient's temperature and record it, to the nearest two-tenths of a degree"--The minimum acceptable standard here is the nearest two-tenths of a degree, not the nearest tenth.
Quality	Subjective quality--performance must achieve certain characteristics which are measured qualitatively (or product must have certain subjective characteristics--for example, boots must have a bright shine).	"Be able to land a UH-1D helicopter with power off using auto-rotation, and making a soft landing from 1,000 feet"--The standard of a "soft landing" is qualitative. Care must be used to define standards of subjective quality as precisely as possible so that two observers would agree in most cases.
Time	Time requirements--performance must be done at a certain minimum speed.	"Correctly multiply pairs of five-digit numbers using a desk-top calculator. The trainee must be able to get the correct answer for at least 10 such multiplications per minute. It is important for this trainee to be able to multiply quickly using this calculator, hence the time requirement. Words-per-minute is a similar requirement for typists.
Time	Production rate--performance must yield a certain daily or monthly output. (Products must be completed at a certain rate.)	"A three-man wire team should be able to lay and splice in three miles of wire per day over moderately difficult terrain, connecting at least three different locations"--Here the amount of wire laid per day, rather than a certain minimum speed, is what is important.

Figure 2-2. Six Types of Standards

Separating Objectives Into Their Three Parts

It is a relatively easy matter to separate an objective into its three parts. Let's look at a couple of examples of doing this.

Consider this objective: "Given a map with two points circled and a protractor, be able to measure the grid azimuth from point A to point B and state the correct answer (within ± 2 degrees) in 120 seconds or less." Here is how you would divide the objective into its three parts:

- Performance. One performance is called for: "Being able to measure grid azimuths." Note that measuring azimuths is the main intent of the objective, while stating the azimuth is the indicator of the performance. You would have no doubt about the trainee's ability to state something; what you want to know is if he can correctly measure grid azimuths (but you'll only know this if he measures it, then states it.) Other indicators might include writing the grid azimuth, checking the correct answer on a multiple choice list of five alternatives, etc.

- Conditions. The conditions stated in this objective are "givens," that is, the map with two points circled and a protractor. Environmental conditions are not important, so they are not stated. You could assume that the trainee would have to be able to perform this task under any ordinary conditions--indoors or outdoors, in bright light or relatively dim light, etc.

- Standards. Two standards are stated in this objective. First, the trainee must state the correct grid azimuth within ± 2 degrees. This is a "minimum acceptable level" standard. Second, the trainee must perform the task within 2 minutes. This is a "time requirement" standard.

Now consider this objective: "Using an M543 wrecker and an M-6 sling, the wrecker operator trainee will be able to operate the hoist as directed in unpackaging the Honest John Warhead section following the sequence specified in TM 9-1340-202-12. Performance will occur on an outdoor, flat, hard surface."

Dividing this objective into its parts:

- Performance. Operating the hoist. Here the main intent of the objective can be directly observed and needs no indicator.

- Conditions. There are several conditions stated throughout this objective; conditions are not clustered in one part of the objective. First, the equipment to be used is specified. Second, the material to be operated on (the warhead) is specified. Third, the environmental conditions are described. And finally, special instructions are implied: the trainee will be directed in his operation of the hoist. So, in this objective, all four types of condition statements (what student has available to work with, what he is to work on, environmental circumstances, and limitations/ special instructions) are used.

- Standards. In this objective, the standard is of the standard operating procedure type. In order to satisfy the objective, the trainee must follow the sequence specified in the appropriate technical manual for the Honest John Rocket System. All steps in the sequence must be completed. No time standard is suggested in the objective, but you could infer that the task must be performed within reasonable time limits.

As you have seen objectives may or may not be "neatly packaged." That is, you may have to dig a little to find the performance required and to organize the conditions and standards that apply, and express them in terms of performances which can be observed. To be suitable for use in developing test items, an objective must contain explicit statements of performance, conditions and standards. If it doesn't, it won't be much help to you.

Just having the essential three parts, however, doesn't automatically make an objective suitable for test development purposes. Objectives can have all three parts and still be inadequate.

ASSESSING THE ADEQUACY OF THE OBJECTIVES

There are four major checks that you should make in assessing the adequacy of objectives. These checks will be facilitated by working from your list of objectives broken down into their three parts (performances, conditions and standards). The checks include determining that:

- Each objective covers a single task, and is not a combination of tasks.
- The main intents of objectives are clear.
- Performance indicators are simple, direct, and part of what the trainees can already do.
- Performances, conditions and standards are specified in precise, operational terms.

Figure 2-3, a foldout at the end of this chapter, shows the sequence of operations for checking the adequacy of your objectives. We will discuss each type of check separately. Please fold out Figure 2-3 at this time.

Checking That Objectives are Unitary

Looking at Figure 2-3, you can see that if any objective given as input to the CRT development process is lacking one or more of its main parts--performance, conditions or standards--you cannot begin to assess its adequacy. Instead, you must send such incomplete objectives back through channels and request clarification. If you think you can fill in the missing parts of such objectives, you may do so, but send them back for approval. When you have received clarification from the originators of the objective(s), you can begin to assess their adequacy.

It is important that the objectives you use to develop a test are unitary--that each covers one task only. It is much more difficult to write test items for compound objectives--those covering more than one task. Figure 2-3 shows that if your objectives each cover only one task, you can proceed to the next step of assessing their adequacy. However, any compound objectives must first be broken down into unitary objectives before proceeding.

To check that objectives are unitary, you should examine the parts that describe the performance. (Remember, this may be labeled as "task," "action," etc.) So, looking at the performances called for in your objectives, ask yourself the following questions:

- Does each objective call for performance on just one task?
- Are all tasks independent? That is, successful performance on one objective does not require successful performance on a preceding one.

If your answer to either question is a definite "no," your objectives are probably not unitary, and need to be broken down into unitary ones. Do this by carefully subdividing them as appropriate. Be sure to seek verification, though, by submitting your list of unitary objectives through channels to their originator.

Remember, when subdividing compound objectives into unitary ones, all that is broken down is the "task" (performance) part of the compound objective. Each unitary objective will include the same conditions and standards as specified in the compound objective from which it was derived.

Let's look at a couple of examples. First, here are the performance parts of three objectives, each of which appear to be unitary:

1. Perform activities for maintenance of the SP Howitzer as specified in the operations and maintenance manual. . .

2. Perform the appropriate before-firing duties for the SP Howitzer as specified. . .

3. Perform the necessary before-operation service activities on the SP Howitzer as specified. . .

Note that each of these objectives covers a single, separate task: (1) maintenance task, (2) set-up task, and (3) service task. Each task is relatively independent of the others. Consequently, there is no need to break these objectives down any further.

Now consider the following objectives which read in part:

1. Treat for shock. . .

2. Treat for nerve gas inhalation. . .

3. Administer mouth to mouth resuscitation. . .

4. Control arterial bleeding. . .

5. Give first aid for burns; chest wounds; abdominal wounds; head, face, and neck wounds; and open arm and open leg fractures. . .

6. Correctly apply a tourniquet and construct a hasty litter.

Note that objectives five and six call for performance on several different tasks, while the other objectives concern single tasks. In addition, there is a lot of overlap--lack of independence--among objectives: For example, controlling arterial bleeding is a part of what must be done in objective five, while treating for shock is probably common to all objectives.

If one were to try to make the above six objectives unitary, it might be done as follows:

1. Treat for nerve gas inhalation. . .

2. Give first aid for burns. . .

3. Give first aid for chest wounds. . .

4. Give first aid for abdominal wounds. . .

5. Give first aid for head, face and neck wounds. . .

6. Treat open arm and open leg fractures (bleeding cannot be controlled by direct pressure, digital pressure to pressure points, or elevation).

7. Construct a hasty litter.

8. Administer mouth-to-mouth resuscitation.

Now the objectives are more nearly independent and cover separate, single tasks. Note that applying a tourniquet is incorporated in objective six--it is not really a separate task, it is a normal part of treating compound fractures where blood flow cannot be otherwise controlled. Also note that objectives five and six may each seem to cover several tasks. They really do not: first aid for head, face, and neck wounds is one task--procedures don't differ. The procedures for treating open arm and open leg fractures are also the same. All tasks covered in the original six objectives are now covered in a unitary fashion by the eight new objectives. No performances have been changed--only broken down into unitary performances. The conditions and standards for each objective will remain the same.

Checking for Clarity of Main Intents

The next check is to ensure that the main intent of the objective is clear. To do this, look at your performance statement for the objective. Then ask yourself:

- "Does the performance statement call for that performance which demonstrates the objective?"

If you can answer this question affirmatively, the main intent of your objective is clear. If your answer is "no," perhaps the performance called for misses the main intent of the objective, or possibly does not provide directly observable performance. In either case, you should make sure that the main intent itself is clear and is defined operationally.

Here are some examples of performance statements in which the main intent is a clearly specified, directly observable performance.

- "Cross a wire obstacle. . ." The performance called for is crossing a wire obstacle and that is the main intent. Crossing the wire can be directly observed.

- "Unlock the security container. . ." Unlocking is directly observable, and the objective's main intent is that a person be able to unlock the container.

Here is an example of a performance statement in which the main intent is clear but the performance called for is an indicator:

- "Circle the picture of the proper shears to use for cutting a curved line in sheet metal. . ."

Circling the picture is the performance called for, but certainly not the main intent of the objective. The main intent is clear, though--knowing which type of shears to use for the task. If the objective wanted the individual to know which type of shears to use and how to use them, it might have been stated as follows:

- "Given five different types of shears, select the proper shears and cut a curved line in the piece of sheet metal." In this case the main intent of the performance is cutting a curved line with the appropriate shears; there is no indicator.

The following are examples of performance statements in which the main intent is unclear and no indicator is provided:

- "Be aware of techniques for setting up a drop zone. . ."

"Being aware" of something is vague and ambiguous. How could a trainee show that he is "aware"? What action is called for? Does the objective want the person to be able to set up a drop zone, or supervise setting up, or teach how to set up a drop zone? You can't tell from the performance statement because the main intent is unclear. Also note that there is no indicator provided which would tell you how to measure "being aware."

- "Demonstrate an understanding of the differences between treating a simple fracture and a compound fracture. . ."

As in the preceding example, the main intent is unclear; you don't really know the purpose of the objective. Are you supposed to find out if an individual can treat both types of fracture, or are you supposed to see if a person tries to treat a compound fracture like a simple one? You can't tell. Also there is no indicator to help you figure out how you are supposed to measure the "demonstration of an understanding." So you really don't have any idea of what performance is called for, though at first glance the statement may have appeared to actually state a performance.

Finally, let's look at some examples of performance statements with clear indicators but with unclear main intents.

> It is important to know what the main intent is, even when there is a clear indicator, otherwise you can't know whether the indicator is really acceptable because you don't know what it is supposed to indicate.

Consider this example:

- "Place a check mark beside the part numbers of the parts needed to replace the brush assemblies on the 45 KW generator. . ."

Note that the indicator is perfectly clear but that the main intent is not readily apparent. The main intent could include any of the following:

- Be able to select the correct parts for replacing generator brushes.
- Be able to correctly read and interpret a list of part numbers.
- Be able to fill out a request for replacement parts.
- Be able to sort parts needed for one repair task from parts needed for another repair task.

So you really don't know what the indicator is supposed to indicate.

Now look at this example:

- "Demonstrate an understanding of good briefing skills by listing the three main parts of a briefing. . ."

Here the indicator is clear; it calls for an observable act--listing. And it might sound like the main intent is clear. But is it really? Does "listing the three main parts of a briefing" demonstrate an understanding of good briefing skills? Listing the main parts of a briefing only indicates an individual's knowledge of such parts, not his ability to conduct a successful briefing nor even to recognize whether a particular briefing is organized in three parts. Although the main intent is stated, it is not clear. In any case, the indicator doesn't even seem to be in the same ballpark. The point is that you don't really know what the main intent is, and the indicator doesn't give you any help in interpreting it. Maybe the indicator is the performance that the person who wrote the objective wants measured and the main intent was just poorly stated. Or perhaps the indicator is poor and the main intent should be clarified and supported by a different indicator.

In summary, the performance statements for any objectives from which you have to develop a test must contain clear main intents. If the intent calls for a performance that is not directly observable, an appropriate indicator must be provided. When you cannot be sure what the main intent of an objective is, it must be revised, clarified and approved before you begin the next check.

Ensuring That Performance Indicators Are Simple, Direct, and Part of the Trainees' Repertoire of Behavior

Figure 2-3 shows that if the main intent of the objective is clear, you must next ask whether it is overt or covert. An overt main intent is one which is observable and measurable. In the preceding section, the examples of "cross a wire obstacle" and "unlock the security container" were overt main intents. Overt main intents do not require indicators: They already tell you what performance is called for and how to measure it.

Covert main intents require indicators since the performances they require are not directly observable. A covert main intent tells you the unobservable performance which the objective is about, while its indicator tells you how to measure whether or not an individual can perform it.

If your objective's main intent is measured through an indicator, you should make sure that the indicator is appropriate. A good indicator is:

- Simple. That is, it is as uncomplicated as possible. You don't want the main intent obscured by an unnecessarily complicated indicator.
- Direct. Indicators are used when the performance called for by the main intent of the performance statement is either not directly observable or not practical in the testing situation. But the indicator should be as straightforward as possible. It should allow you to determine whether or not the main intent has been satisfied without your having to go through chains of inference.
- Part of the trainees' normal repertoire of behavior. The trainee should be able to perform the indicator behavior: The indicator behavior itself is not what you want to train or test. You only use it as a measure of the main intent. So it is important that the indicator is simpler than the main intent and that the trainee can do it. If the indicator were not a part of the trainee's normal repertoire, you would be measuring two things--performance on the indicator and performance on the main intent.

Let's analyze some examples of indicators to see if they are as simple and direct as possible, and part of the normal repertoire. Here's the first example:

> "Show that you can recognize the major bones of the human skeletal system by drawing a picture of each bone beside the names of the bones provided on a mimeographed handout."

Okay, recognizing bones is the main intent, while drawing pictures of bones is how you indicate recognition. Drawing pictures of bones is a direct indicator in this case, since if a person can draw the correct picture next to the name of a bone, you know he can recognize the bone--you don't have to make any inferences. But drawing a picture is not the most simple indicator. Worse yet, drawing a bone well enough so that an examiner could identify it is not a part of the trainees' normal repertoire unless the trainees happen to be skilled illustrators. Thus, a person could fail to satisfy the objective because he can't draw well, not because he can't recognize the bone.

In fact, the indicator is a poor one for another reason: The main intent is to recognize bones but the indicator requires the person to recall what it looks like, then draw it.

A better indicator for this main intent would be ". . .by writing the name of the bone next to the picture of the bone" or, better yet, ". . .by choosing the correct name from the list provided and writing it next to the picture of the bone." (The pictures of the bones are provided on a mimeographed handout.)

Now consider this example:

- "Be able to recognize properly filled-out and improperly completed orders. Show your ability to do this by writing examples of each."

The indicator is "by writing examples of each." This indicator appears to be neither simple nor direct. The performance called for is a complex one--writing orders--and you would have to infer that an individual could recognize properly and improperly filled out orders based on his ability to write examples of each. In addition, the indicator behavior required appears to be more difficult than the behavior that the main intent is concerned with--the ability to discriminate between properly filled out orders and those which have not been properly completed. Thus, the indicator is less likely to be a part of the individual's repertoire than the main intent; this is exactly the opposite of the way things should be.

A better indicator would be ". . .by sorting examples of orders into two piles--those that are properly filled out and those that aren't." In this case, all the individual has to do is sort documents--a simple and direct indicator of ability to recognize proper and improper documents. This indicator would also be well within the normal behavioral repertoires of most trainees.

In summary, if the main intent of an objective is covert--not directly measurable (for whatever reason)--you should check to be sure that an appropriate indicator is included. Such an indicator will be as simple and direct a measure of the main intent as possible, and will require a behavior which the trainee is easily able to perform.

If indicators are not adequate--because they are not simple or direct enough, or not a part of the trainees normal repertoire of behavior--or, if necessary indicators are missing, you may modify the indicators or create new ones. Be sure, though, to have them approved by the objective writer. If you don't feel you can properly modify or create a new indicator, you should request improved indicators. When the necessary indicators are revised and approved, proceed with the final check on the adequacy of your objectives.

Checking That Performances, Conditions, and Standards are Specified in Precise, Operational Terms

The third check you should make for an objective is to ensure that the statements of performance, conditions and standards are written in precise, operational terms. This means that each statement should be easily translatable into actions. You have essentially already done this for the statement of performance by checking for clarity of the main intent and appropriateness of the indicator. A further check on the performance statement of your objective will be helpful at this point, though.

Make sure that the statement of performance uses a specific, action verb and you've about won the battle. Figure 2-4 shows examples of verbs often found in the performance statements of objectives. The left half of Figure 2-4 shows examples of non-action verbs which generally are not suitable for performance statements. The right half shows examples of action verbs which may be suitable. Of course, it is impossible to list all appropriate action verbs or all inappropriate non-action verbs.

Non-Action Verbs	Specific Action Verbs
Appreciate	Brake
Be aware of	Check off
Be familiar with	Label
Feel	Solder
Know	State
Understand	Turn

Figure 2-4. Examples of Verbs Often Used to Specify Performance in Objectives. (Only those on the right are really suitable.)

Sometimes what sounds like an action verb may not be suitable in a particular context, and what appears to be a non-action verb may designate observable actions. So, use Figure 2-4 simply as examples of non-action and action verbs. If the verb in a performance statement is more like those on the left side than those on the right side of the Figure, the performance is probably not stated in terms precise or operational enough for you to use. But always examine the verb in the context of the statement of performance and determine if it is as specific an action verb as possible.

Statements of conditions and standards must also be written in precise, operational terms. If they are not, you will not have enough information to build an adequate test. Figure 2-5 shows examples of statements of conditions and standards, some of which are specified in precise, operational terms, and some of which are not. The column on the left shows what the standards or conditions are supposed to say in certain objectives. The right column shows how such meanings could be incompletely or incorrectly specified. Note that properly specified statements of conditions will tell you all you need to know in order to set up the appropriate conditions for a test. Standards must tell you as precisely as possible how the individual will be scored--about how is not good enough. You, the item writer must actually determine how to comply with the standard when you write an item. For example, if the objective calls for 80% accuracy you must decide whether this means 4 out of 5, 8 out of 10, 16 out of 20, etc.--based upon your assessments of the requirements of the situation, and of the resources available. Also note that, at first glance, some of the poorly specified conditions and standards might appear adequate.

	If The Condition Or Standard Is Intended To State:	Then This Is An Improperly Specified Statement:
Conditions	• Given a 45 KW generator with a broken shaft bearing. . .	• Given a malfunctioning generator. . .
	• . . .under ordinary field conditions in daylight	• . . .under ordinary conditions
	• . . .using a multimeter and signal generator only	• . . .using appropriate test equipment
	• . . .without getting glue on the movable surfaces	• . . .taking proper precautions
Standards	• . . .following the sequence specified in the Field Artillery Rocket Crewman's SMART book	• . . .following the best sequence
	• Using a 10" slide rule, multiply two five-digit, two-decimal place numbers and write the answer to the nearest tenth	• Using a slide rule, multiply two five-digit, two-decimal place numbers and record the correct answer
	• . . .typing at least 60 words per minute corrected for errors	• . . .typing at a quick rate
	• . . .the steak should be light to medium pink in the middle	• . . .the steak should be of an acceptable color in the middle

Figure 2-5. Examples of Statements of Conditions and Standards

You should ask yourself, "Does it really tell me all I need to know to establish proper conditions or proper standards, or will I have to supply information on standards and conditions myself?" If your answer is that you'll have to supply information or fill in details, etc., then the conditions and standards are not specified in precise, operational terms and you won't be able to use them. If you tried to use them, you'd run the risk of going through a lot of effort and ending up with a useless test.

Note that appropriate conditions and standards are often related to the level of your objective--that is, at what level the objective specifies performance. For example, a Level One objective may be to repair any malfunctioning generator. In this case "given a malfunctioning generator" is an appropriate statement of conditions. If however, the objective is to repair a 45 KW generator with a broken shaft bearing--any malfunctioning generator may not conform to these requirements, and therefore would be an inappropriate condition.

When you review objectives, if you find some that do not have tasks, conditions, or standards specified in operational, precise terms, you should not proceed with test development activities. Instead, send each inadequate objective back to its originator. You should attach a sheet to each inadequate objective spelling out what is wrong with it, and why you cannot develop a test for it until you receive clarification. (Be sure you are not nit-picking and that the objective really doesn't give you enough information.) Then, wait until you receive such clarification before you begin the next step of test development.

Summary

Let us review what you have done so far. Up to this point you have examined the three parts of your input objectives, made sure that all objectives are unitary, ensured that their main intents are clear and that appropriate indicators are used when necessary, and have checked to see that all parts of the objectives are specified in precise, operational terms. Whenever a check has revealed that an objective is inadequate you have either modified it and sent it back for approval, or documented the problem and sent it back for revision. Objectives may have been considered inappropriate for one or more of the following reasons:

- One or more of the objective's three parts were missing
- An objective covered more than one separate task
- Main intent was unclear
- Indicator was improper
- Performances, conditions or standards were not specified in precise, operational terms.

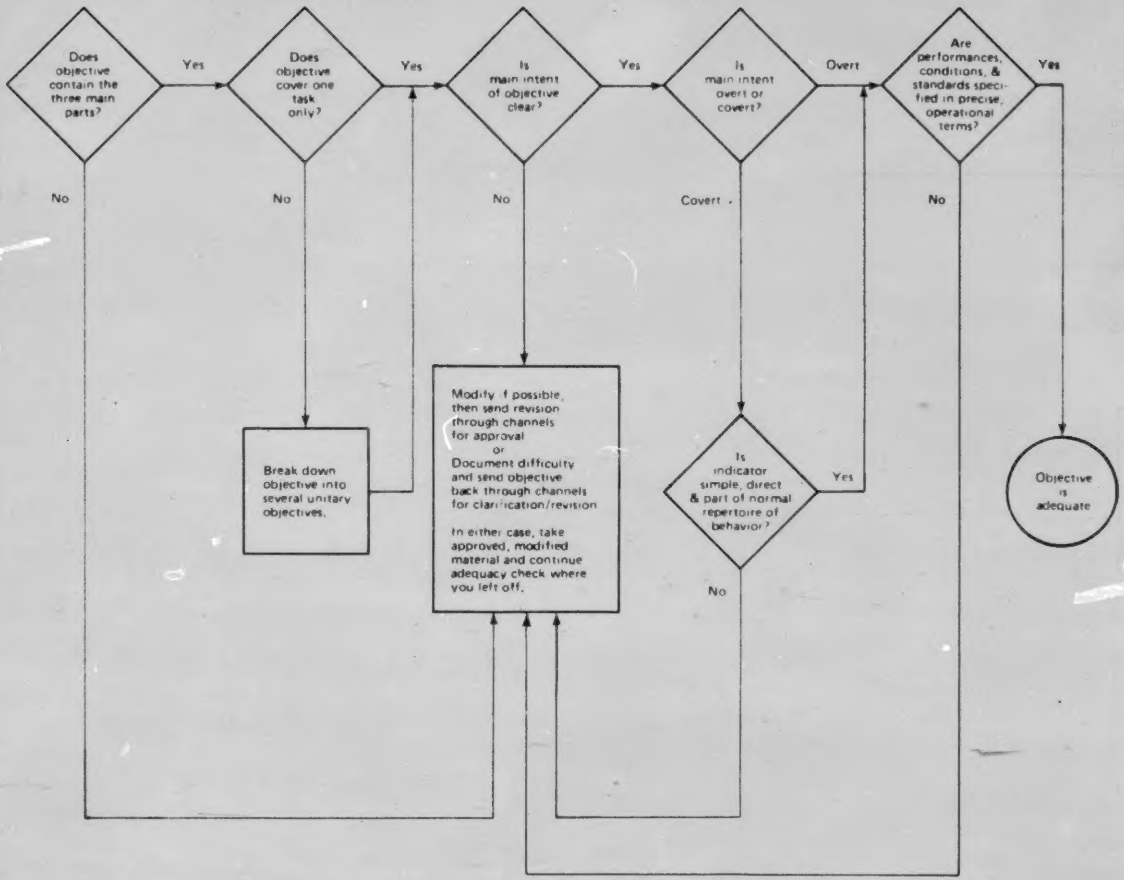

Does objective contain the three main parts?
Yes
Does objective cover one task only?
Yes
Is main intent of objective clear?
Yes
Is main intent overt or covert?
Overt
Are performances, conditions, & standards specified in precise, operational terms?
Yes
No
No
No
Covert
No
Break down objective into several unitary objectives.
Modify if possible, then send revision through channels for approval or Document difficulty and send objective back through channels for clarification/revision
In either case, take approved, modified material and continue adequacy check where you left off.
Is indicator simple, direct & part of normal repertoire of behavior?
Yes
No
Objective is adequate

Figure 2-3. Sequence of Operations for Assessing the Adequacy of an Objective

CHAPTER 3

DEVELOPING A TEST PLAN

Now that you've assessed the adequacy of the objectives on which your CRT will be based, and modified them as necessary, you are ready to plan the test itself. Developing a test plan is an important step in CRT construction. In this step, you consider factors which will enable you to construct test items based upon objectives.

Figure 3-1, which folds out at the end of this chapter, shows the sequence of operations involved in developing a test plan. Please fold out Figure 3-1. First, you examine practical constraints, such as time and manpower availability, to determine if they affect how the objectives are to be tested. Then, if such constraints are problems, you must decide how to proceed--either by developing a plan for selecting among objectives or, if that is not workable, by modifying objectives. Next, you plan the type of items (item format) to use in the test, and their level of fidelity (realism). Then, if necessary, you develop plans for item sampling and for sampling among conditions. Finally, decide how many items should be included on the test, and document the entire test plan. You then can use this plan to guide you in constructing a pool of items--which is covered in the next chapter.

EXAMINING PRACTICAL CONSTRAINTS

Now that you have checked your objectives closely to make sure they are adequate, you must examine them to see that they are actually administrable. To do this you need to take into account several different types of practical constraints by gathering as much information as possible on test administration and training conditions. Practical constraints include:

- Time availability
- Manpower availability
- Costs
- Equipment and facility availability
- Degree of realism in training and degree of realism required in testing
- And others

Note that these types of constraints are all interrelated. For example, time availability, manpower availability, equipment availability, and costs are often all different aspects of the same problem.

Time

The first type of practical constraint, time availability, is easily understood. Often the situation is such that it is impractical to test the objective as it is stated in the available time. Perhaps the objective is "March 25 miles through marshy terrain during inclement weather conditions in 12 hours" or "Watch a radar scope for enemy blips for 14 hours, maintaining proper vigilance as indicated by detecting the three simulated enemy blips presented during the interval." Both of these examples would take much too long to test practicably in most situations. These objectives may have to be modified to permit testing in less time. In general, time limits must be placed on test administration, which in turn may limit the objectives being tested. Sometimes there are several objectives which, if tested, would take more time than is available. In such cases, it may be possible to select among these objectives without having to modify them.

Manpower

Manpower availability can also impose practical constraints. For example, if under normal conditions, it takes 4 men to operate a main battle tank--a commander, driver, gunner, and a crewman/loader--and you want to test a class of assistant crewman/loaders under normal operational conditions, then personnel trained in the functions of commander, gunner, and driver will be required for the test. If these personnel are not available, then there is insufficient manpower available for conducting the assistant crewman/gunner test under normal operating conditions.

Often manpower constraints are severe when only a few test administrators will be available, yet many trainees will have to be tested concurrently. For example, 20 soldiers may be tested simultaneously in basic first aid procedures by only two administrators who must thus try to monitor the performance of ten individuals apiece. There are many instances in which an objective appears to call for more manpower than is available. In such instances you may wish to select among objectives, so that enough manpower can be available for the testing, or to modify objectives so that less manpower is required.

Costs

Cost is an important factor in developing CRTs. The cost of test administration must be kept within the limits dictated by the testing budget of the facility where the test will be used. For example, it would be entirely too costly (and unreasonable) to have a demolition-specialist trainee blow up a bridge to test his ability to achieve maximum damage. There must be other more practical means of testing this objective. If the actual objective specifies demolishing a bridge, it may well have to be modified so that the bridge is not actually demolished, but the trainees demonstrate the processes leading up to demolition.

If the cost of testing all objectives is prohibitive, and if selecting among objectives is feasible, then the best alternative may be to test a subset of them.

Facilities/Equipment

Often the situation is such, that equipment and/or facilities are not available for test administration. This is especially true for sophisticated equipment and very specialized facilities. For example, how can a trainee demonstrate competence in escape and evasion in a tropical jungle, when the testing must take place in the Southwestern United States? An extreme example of a facility-caused constraint may be firing a missile down range. At many test sites it is impossible to obtain a range that is long enough.

An example of a practical constraint concerning equipment availability might involve a course on troubleshooting a terrain-following radar system. The performance objective may include planting a bug in the system and having trainees locate the problem and replace or repair the necessary parts. However, this radar system is sufficiently complex and costly that it is not made available for training purposes and therefore prohibits testing on the actual equipment. In this case, equipment availability is a very severe practical constraint. Another example is troubleshooting a computer: The downtime of the computer may be so costly as to negate its use for training purposes.

If you have many objectives which would tax facilities/equipment beyond feasible limits, it may be possible to select among them rather than to modify the objectives.

Degree of Realism

Another important practical constraint that may impact on CRT development is establishment of an acceptable degree of realism in training and testing. Consider training in first aid: In almost all cases of teaching first aid for an open leg wound, a patient with such a wound is not available even for observation, let alone practice. A suitable substitute must be made here, thereby decreasing the degree of realism. Another such case, just as obvious, is training disarmament of live mines. The mines, of course, in training are never live; therefore, the training conditions are not very close to the real situation.

A high degree of realism in testing is also similarly difficult to provide. In testing basic marching maneuvers associated with the drill and ceremony component of basic training, a parade field is necessary. The degree of realism in testing decreases as the dimensions of the testing field differ from a standard parade-size field. How real are the testing conditions if a 40-ft field is being used? Another example involves testing trainees on detecting and challenging intruders. How real is a testing situation where the test administrator jumps out at a trainee while the other trainees wait within hearing distance for their turn? The degree of realism should not differ from training to testing.

There are other practical constraints which you may encounter in the development of your test; however, this section has covered the most common ones. Less common types of practical constraints include:

- Logistics
- Supervisory effectiveness
- Communications
- Ethical considerations
- Legal considerations

Remember that in most cases constraints are interrelated. As you'll recall, the practical constraint in the example of the terrain-following radar system was categorized under equipment availability. This constraint could also be categorized under costs. Another instance of interrelation was in the example of the 40-ft. field being used for testing basic marching maneuvers. Not only was the degree of realism low, as indicated in the example, but the objective was limited by facility availability.

Potential Sources of Data

Information on practical constraints can be obtained from a variety of sources. One source is current documentation on test administration and training conditions (such as Army Field Manual 21-6, TRADOC Reg 350-100-1, TRADOC Pam 600-11, etc.). These documents are good sources for current procedures in this area, but more direct sources of information on training/testing situations at specific locations are preferable. Such direct sources include personal experience and observations, and the observations of your associates, especially those who have given similar tests before at the same place. The best single source of information on practical constraints at a particular site, is a visit to that site. If possible, you should arrange to go to the site and observe first-hand the availability of facilities, equipment, and manpower. While there you should talk with personnel who conduct training and testing to find out more about time availability and budgeting considerations at the site.

Other sources of data may also be available to you. Use your discretion and ensure that this information is accurate.

Assessing Practical Constraints

After you have identified practical constraints, you must determine whether they are severe enough to prohibit testing all objectives as stated. As you have probably noticed, some constraints may be very strong, while others are relatively unimportant. Each must be considered carefully. Some constraints may be so severe that they necessitate modification of the objectives, or selecting among objectives, whereas other constraints may be easily overcome.

As you can see from Figure 3-1, if practical constraints do not constrain testing of all objectives as they are stated, there is no need to either select among objectives or modify objectives.

However, if practical constraints prevent testing of all objectives as stated, you will have to select among objectives or modify objectives. First determine whether it is feasible to select among objectives. It often is feasible, unless objectives concern critical tasks.

> When your objectives concern critical tasks, you should probably not select among them. That is, if misperformance could lead to loss of life, property, or mission failure, you should be sure that everyone can meet every objective.

Then determine if selecting among objectives will overcome practical constraints. Sometimes selection won't overcome practical constraints since it is possible that any one objective, as stated, would overtax resources. So, before deciding to select among objectives, make sure that doing so will solve the constraints problem.

Selecting Among Objectives

If it is feasible to select among objectives, and doing so will overcome practical constraints, then, instead of modifying objectives--which runs the risk of distorting their original intent--you include objectives as originally stated, by selecting among them. Don't inform trainees which objectives you intend to test, however. If trainees know they may be tested on any objective, but don't know which, they must prepare for all of them. Let's look at an example.

Suppose we are developing a CRT to use in evaluating pie-making-ability in a food service course. Assume that there are 10 testable objectives. Each involves being able to bake a pie which is rated as adequate by three independent judges. The following 10 pies are taught:

- Apple pie
- Cherry pie
- Peach pie
- Blueberry pie
- Coconut cream pie
- Pecan pie
- Raisin pie
- Black raspberry pie
- Banana cream pie
- Lemon meringue pie

Now, assume that the training lasts 10 hours (1 hour per pie) and that 100 students are to be tested. We have two hours available for our end-of-unit CRT. It is prohibitively expensive to provide sufficient ingredients for each student to bake each pie. Here is a case where we might legitimately select among objectives in developing CRTs, rather than testing on each individual objective. Thus, trainees might be tested on their ability to prepare only two pies (one fruit type and one cream type). This is an example of "stratified" selection among objectives. One objective is selected from each of two strata. If all pies were of the same type, there would be no strata, and any objective could be randomly selected.

Similarly, if an electronics repairman was to be tested on his ability to fix radios, oscilloscopes, and signal generators; objectives might be selected randomly from among these three strata. Thus, he would be tested on repairing at least one radio, one oscilloscope, and one signal generator.

In any case, it is important that the trainees not know which particular objectives (which pie, which radio, etc.) they will be tested on. They must be responsible for all objectives.

Two important aspects of selecting among objectives in CRT development are indicated in Figure 3-2.

When selecting among objectives in CRT development be sure that:

- The objective or objectives to be tested are chosen at random from the entire population of objectives available for testing
- The students to be tested are not informed of the sample of items selected for testing

Figure 3-2: Guideline for Selecting Among Objectives in CRT Development

Remember, if you select among objectives, you can only guarantee that trainees can perform objectives on which they were tested (and passed). You can also document the testing procedure to inform people that trainees were responsible for all objectives, did not know which they would be tested on, and had an equal chance to be tested on any objective (since you selected at random from among the objectives). As noted, this is not appropriate for critical objectives, but it will be satisfactory for many others.

Document your plan for selecting among objectives so that you will have a record of how to do it when you build your test. Documentation might simply say: "Select randomly any two of the five objectives," or (as in the case of the pie-making example), "Select any one fruit pie randomly, and any one cream pie randomly."

Modifying Objectives in Light of Practical Constraints

In light of the constraints found, objectives may have to be modified. Consider the three parts of objectives discussed earlier: performances, standards and conditions. Performances should not be modified unless absolutely necessary. Standards, on the other hand, may be modified. For example, you may have to lengthen or shorten time limits for testing. In many cases you will find it necessary to modify conditions, such as settings, locations, etc. Assess each constraint separately and modify the objective as required. Modify as little as possible to make the objective acceptable and accurate, but still appropriate for testing.

Now let's look at an example of a situation in which you would have to modify an objective because of practical constraints in the training/testing situation. Here is the objective:

"Given a complete field kitchen set-up, the basic cook trainee will prepare a standard dinner meal for 250 persons under tactical forward area mess conditions. The meal must be prepared within 3 hours, and the student must follow hygienic regulations as specified in the POI for Basic Cook. The trainee will have a food service apprentice under his supervision. Food will have to be prepared with a minimum of noise and light, and normal perimeter security regulations must be observed. The meal must be rated as satisfactory by three judges all of whom have held the MOS for Basic Cook for five years and have been first cook for at least three years."

You make a site inspection of the facilities where the testing is to be conducted and find the following facts which you feel are potential practical constraints:

1. A test range equivalent to a forward area is not available.
2. An average of 14-16 men are trained at once for the basic cook MOS. Total test time available for the field kitchen unit is 12 hours and must include tests of setting up the field kitchen, maintaining equipment and preparing morning and afternoon meals.
3. The training budget will not allow for food for feeding 250 people per test--food cannot be wasted. All food prepared must be eaten according to the SOP at this facility.
4. Three cooks, each with three years experience as first cook, are not available for testing purposes. Only one such individual is available. There are several other cooks available, but none has served as first cook.
5. Only three test administrators are available.
6. Only two field kitchen set-ups are available.

Considering the above information on practical constraints, it should be obvious that the objective must be modified before a test can be developed which will be suitable for that facility. The question is "how can the objective be modified so as to not violate its intent?" Let's consider the types of constraints and analyze how they affect the objective.

First. facility and equipment constraints do not appear important: There are two field kitchens available which should be ample. Although there is no test range similar to a tactical forward area, such an area can be simulated. The simulation can be made more realistic by playing tape-recorded "field" sounds (artillery, fire bursts, etc.), requiring minimal cooking sounds, and maintaining minimum lighting. The resulting loss in fidelity should not be critical in this situation.

Manpower constraints do appear serious, though, on several counts. With only three test administrators, it will be hard to determine whether trainees are following specified hygienic regulations. Another manpower constraint has to do with the three cook/judges specified in the "standards" portion of the objective--only one such cook is available to participate. The "judges manpower constraint" can be disposed of now: the objective's specifications for judges are probably too rigorous. They can be relaxed without seriously affecting the intent of the objective (measuring the trainees' ability to prepare a satisfactory meal). The objective could be easily modified to read "...rated as satisfactory by three judges currently holding the MOS for basic cook and all having at least six months experience." This is a much lower requirement for the judges, but should be appropriate and adequate for the test situation.

Time constraints are quite severe. Assuming that the other tests which must be given for the field kitchen unit (setting-up, maintenance, etc.) will require two-thirds of the 12 hours available, only four hours are available for testing 14-16 men--and each must be tested on his ability to prepare a meal for 250 people within three hours. Obviously, the time constraints are too severe to get around by trying to stretch time availability for testing or by slightly lessening the time requirements stated in the objective. But, since time constraints are interrelated with manpower availability, they can be overcome by manipulating the manpower.

Given the two field kitchen setups available, two groups of trainees can be tested at once. Although the objective specified the trainee being tested with a food service apprentice to help him, it should not alter the spirit of the objective to require the trainee to serve either as supervisor or as food service apprentice. If we modify the objective in light of this, we can now test two teams of two trainees (one supervisor and one apprentice)--one at each field kitchen setup.

Now, the requirement that a meal be prepared for 250 troops is probably over-stringent. The trainee could just as easily demonstrate his ability to prepare meals for large groups by preparing a meal for 100 troops. This should take only about two hours instead of three. If we modify the objective accordingly, we can now have two teams of two working concurrently at each field kitchen. Thus, 16 trainees can be tested in four hours.

All trainees can take a brief written test on planning evening meals for 250 troops--quantities of supplies involved, scheduling, logistics, etc.--and on managing food service assistants. Thus, whether a trainee served as cook (supervisor) or apprentice, he would be tested on planning and managing preparation of an evening meal for 250.

Finally, there is a cost constraint: food cannot be wasted. This is not an important constraint, since it can be easily overcome. A total of 800 troops could be fed from the meals produced by the eight groups of trainees. These 800 portions could be served to other troops on field exercises in the area, if scheduling were coordinated. Alternatively, the prepared food could be trucked to a mess hall and served as the dinner meal.

It is helpful to make a table of the conditions and standards in an objective that requires modification in light of practical constraints. Figure 3-3 shows such a table filled in with information from the food service example we have been discussing. Note that the table presents the conditions and standards which require change, why they require change, and how they should be changed.

Use of a tabular summary such as Figure 3-3 will help you organize information on modifying objectives to overcome practical constraints. By using a summary table, you won't lose sight of the forest by concentrating on the trees.

Here is how the objective might read after modified by practical constraints:

> "Given a complete field kitchen set-up, the basic cook trainee will help prepare a standard dinner meal for 100 troops under simulated tactical forward area mess conditions. The trainee may serve as cook or food service apprentice. A team of one apprentice and one cook will prepare the meal within two hours. The food will be prepared with a minimum of noise and light and normal perimeter security regulations will be observed. Proper hygienic practices, as specified in the POI for Basic Cook, will be followed. The meal must be rated as satisfactory by three judges currently holding the MOS for basic cook and all having at least six months experience. In addition, the meal must be suitable for consumption, as specified by standard food service regulations, since it may be served to actual troops."

Submit Modified Objectives

After modification, send the objectives back to their originator for approval before proceeding. Be sure to include reasons for modification with the modified objectives. By doing this, you make sure that the modified objectives are suitable--that modification has not distorted the original intent of the objectives.

Conditions and/or Standards Which Require Change	Why These Conditions and Standards Require Change	How to Modify Conditions and Standards so they Overcome Practical Constraints
"250" people must be fed	Can only cook for a maximum of 100 people, not 250	1. No modification, because procedures don't change significantly when going from 100 to 250 people
	Planning a meal for 100 people is less involved --in terms of supplies, scheduling, assistance required, etc.--than planning a meal for 250 people.	2. Take paper and pencil test: estimate amount of food and utensils for 250 3. Indicate how assistants would be managed
3 master cooks each with 3 years experience	Manpower availability: cannot get three highly experienced cooks	Substitute less experienced cooks to do the routine aspects of the judging.
"Supervise one apprentice"	Manpower availability	Have one trainee serve as an apprentice.
"Location in forward tactical area"	Availability of equipment & facilities: Forward tactical area not available	Simulate Forward Tactical Area: 1. Play tape recorded "field" sounds: artillery, etc. 2. Maintain minimum lighting, minimal cooking sounds
"3 hour time limit"	Too many trainees to devote 3 hours to test each one	1. Test two at a time for about 2 hours each (feasible, if meal is for about 100 people) 2. Have one trainee serve as an apprentice

Figure 3-3. Tabular Form for Summarizing Conditions and Standards that Require Change in an Objective and How to Change Them. (With Sample Information from Food Service Example)

PLANNING ITEM FORMAT AND LEVEL OF FIDELITY

Before constructing your test items, you will be faced with questions of item format. Do you want:

- Paper and pencil items?
- Hands-on performance items?
- Multiple choice items?
- Recall measures?
- Job simulations?
- Supervisor or peer ratings?

Virtually any of these formats can be adapted to any testing situation. There may even be others that are more appropriate. Which should you choose? These are questions involving item format and test fidelity.

First, let us discuss what we mean by the term "fidelity." The term "test fidelity" addresses the extent to which a CRT resembles the actual objective (or performance) being tested. The more the CRT resembles the performance in question, the higher the fidelity of the CRT. It is probably obvious to you that this is one place where practical testing constraints have a direct impact on CRT development. If, for example, it is too costly to use an actual aircraft for a maintenance test and you must therefore use a simulator, you lose fidelity--unless the simulator is very much like the actual aircraft in terms of required performances. To the extent that the performances required on the simulator approach those required on the actual equipment, the fidelity loss is minimized. Some simulators, however, cause a great loss in fidelity. For example, if the simulator is a series of 35mm slides of an azimuth cursor and the performance required of the trainee is to check which of four alternative slides is most like the required cursor placement, the fidelity loss from an actual operational radar scope is dramatic. One useful test fidelity scale is shown in Figure 3-4.

Fidelity Level		Types of Measurement
Low Fidelity	1	Ask for Opinions
	2	Ask for Attitudes
	3	Measure Knowledge
	4	Measure Related Behavior
	5	Measure Simulated Behavior
High Fidelity	6	Measure "Real Life" Behavior

Figure 3-4. Fidelity Levels and Types of Measurement

Now that you have an idea of what is meant by the term "fidelity," you can see that item format and test fidelity are closely related. Practical testing constraints may dictate the use of a four-alternative multiple choice paper-and-pencil test, for example, because such tests are simple to administer and easy to score, although the test fidelity may be low.

A good guideline for item format is:

> Select the format that best approximates the behavior specified by the objective.

If the instruction is aimed at problem-solving, then the items should address problem-solving tasks and not, for example, knowledge about the required background content. If the instruction is intended to teach how to evaluate a particular performance, the items should be about evaluating that performance, not actually doing that performance.

Item format and test fidelity are difficult issues. Follow the guideline in the box above to the extent possible, consistent with practical constraints. Use a format which will permit the highest level of fidelity practicable.

Basically, it is easier to develop high fidelity CRTs for hard skill subject matter areas (such as electronic maintenance and artillery fire direction computer) than for soft skill areas (such as leadership and tactics). This is because hard skill areas generally include objectives which are more easily specified in terms of concrete behaviors.

Types of Items for Written Tests

Some objectives can best be tested by paper-and-pencil items. Such tests are usually printed on a form with spaces for answers. Paper-and-pencil items are best suited for evaluating knowledge, ability to use information, problem-solving, and written computations. They are sometimes used as low fidelity measures of hands-on performance skills.

Written test items' main advantage is that they can often be easily scored (indeed, in some cases they can be computer-scored) in contrast to performance test items where scoring depends on the test administrator's observations. Therefore, written items are often relatively reliable measures--that is, they measure approximately the same thing each time they are administered. Performance test items, while often less reliable, are usually more demonstrably valid measures--that is, they are more likely to measure what they are supposed to measure. Written items should be used in performance testing only when the performance itself involves writing or when practical constraints (such as time availability) prevent selecting among objectives.

There are several different types of formats which are often used for written test items, including:

- Multiple-Choice Items
- Matching Items
- Completion Items
- True-False Items
- Production Items

Multiple-Choice items can be adapted to almost all types of written tests. The standard best answer (but not necessarily the only correct answer) is included in the test item itself. This type of item is versatile, can take a variety of different forms and can be used to test different aspects of knowledge.

Matching items generally employ two columns of elements. The student is typically asked to match one element from the first list to the most closely related element in the second list. It is preferable to have different numbers of elements in the lists to discourage the student from using a process of elimination when he gets down to the last match.

Completion items may come in two different forms: One being a question that requires a short-phrase answer and the other having one or more internal blanks that require single words or short phrases. You should use care in writing this second type of completion item--too many blanks may make a sentence incomprehensible.

True-false items have many disadvantages:

- Many times such items are built around sentences which are lifted verbatim from training materials (perhaps only changing one word), which encourages memorization.
- Often it is difficult to determine whether items are true or false when the sentences are out of context.
- High scores can be obtained by mere good-guessing since there are only two possible answers.

A good rule of thumb is to avoid true-false test items entirely.

Production items ("essay" items or oral exams) should also be avoided due to their subjectivity. There are many ways a student can express an answer to this type of item which makes scoring extremely difficult. What's worse, an individual who can express himself well in writing or orally has an edge over the individual who cannot, regardless of their relative achievement on the subject matter.

Some general advantages of using written tests include:

- Easy and reliable administration.
- Easy scoring by hand or machine.
- Coverage of a large quantity of material in a relatively short amount of time.
- Easy maintenance of efficient records.

However, it is often hard to relate written tests to job performance. In many cases the student may be able to pass a written test on a performance and not be able to actually perform the required task. (For example, if an individual could pass a written, multiple-choice test on bomb disposal procedures, would you be willing to send him out to defuse an actual, live bomb?) When using an objective written test you should be certain that the test items are suitable for assessing the achievement of the objective.

> Written tests are most often appropriate for testing abstract concepts and objectives which require knowledge instead of performance.

Items For Performance Tests: Process and Product Measures

Performance tests require the student to perform an overt action or series of actions, rather than to verbalize or write (unless the required performance *is* speaking or writing). Figure 3-5 shows a comparison between performance test items and written test items.

WRITTEN TEST ITEMS	PERFORMANCE TEST ITEMS
Primarily abstract or verbal.	Primarily non-verbal.
Items address knowledge and content.	Items are skills, performances or job related decisions.
Items usually address independent aspects.	Items may be sequentially presented. Errors early in the sequence may affect later items.

Figure 3-5. Some Common Differences Between Performance Test Items and Written Test Items

In a performance test, the student actually performs a task and is judged against predetermined criteria. A performance test may involve product measurement, process measurement or both. Before considering types of performance items, let's discuss the problem of whether the items should measure processes or products.

In developing your test plan you will have to determine whether the objectives require measurement of a product (that is, something which is tangible and which can be readily measured as to its presence or absence) or a process (for example, the degree to which a student follows procedures correctly, regardless of the outcome of his actions).

Product measurement is always appropriate if the objective specifies a product. If a product measure is called for, it should be incorporated into the training objective and it should be carried over into the test items. Product measurement is appropriate when:

- The objective specifies a product.
- The product can be measured as to either presence or characteristics (such as voltage, length, etc.).
- The procedure leading to the product can vary without affecting the product.

Process measurement is indicated when the objective specifies a sequence of performances which can be observed, and when the performance is as important as the product. Process measurement is also appropriate where the product cannot be distinguished from the process or where the product cannot be measured for safety or other constraining reasons. Generally speaking, process measurement appears appropriate when:

- Diagnostic information is desired.
- Additional scores are needed on a particular task.
- There is no product at the end of the process.
- The product always follows from the process, but high costs or other practical constraints prevent measurement of the product.

Following are descriptions of conditions which may call for both product and process measurement:

- Although the product is more important than the processes that led to its completion, there are critical points in the processes which, if misperformed, may cause damage to personnel or equipment.
- The process and product are of similar importance but it cannot be assumed that the product will meet criterion levels just because the process is followed at criterion levels.
- Diagnostic information is needed. By having process measures as well as the product measure, information as to why the product does not meet the criterion can often be obtained. That is, if the product does not meet the criterion, then something which has been done wrong in the process may be discovered.

When both process and product measures are taken for a given objective, scoring must follow the criterion specified in the objective. That is, if the criterion specifies only a product, not a process, than process scores cannot be used to assess achievement of the criterion. This, of course, does not preclude obtaining additional process information where such information is useful in an auxiliary way (for example, as diagnostic information) and is feasible to obtain.

One classification has suggested three types of tasks to illustrate the relative roles of product and process measurement:

1. Tasks where the product is the process.

2. Tasks in which the product always follows from the process.

3. Tasks in which the product may follow from the process.

Relatively few tasks are of the first type. Drill and ceremonies, playing a musical instrument and public speaking are examples. More tasks (such as fixed procedure tasks) are of the second type. In these tasks, if the process is correctly executed, the product follows. For example, if you pack a parachute by following the correct process, the product, a properly packed parachute, will follow.

A large number of tasks are of the third type, where the process appears to have been correctly carried out but the product was not attained. There are at least two reasons why this can happen: Either we were unable to specify fully the necessary and sufficient steps in task performance, or we did not accurately measure them. Rifle firing, for example, illustrates that there is no guarantee of acceptable marksmanship even if all procedures are followed. In this case, process measurement would not adequately substitute for product measurement. So, before using a process measure, ask yourself this question:

- "If I use only a process measure to test a man's achievement on a task, how certain can I be from this process score that he would also be able to achieve the product or outcome of the task?"

If your answer is "I can't be very certain," you'd better add a product measure.

Now, let's look at types of items for performance tests. You will see that these items can be used for process or product measurement.

Types of Items for Performance Tests: Process Rating

When using a rating scale, you should specify the rating a student needs to achieve the performance specified by the objective. For example, a scale from 1 to 6 might be used to rate public speaking ability (See Figure 3-6). Here, 6 is the acceptable standard for achieving the objective, while 1 is the beginning level.

Is poor speaker but speaks without speech impediment					Speaks fluently in a well-modulated voice, is interesting, does not pause inappropriately, etc.
1	2	3	4	5	6
Rating needed for entry into the course					Rating needed to pass criterion test at end of course

Figure 3-6: Sample Numerical Scale for Rating Public Speaking Ability

Such a scale might also be used to assess entering behavior at the start of instruction. For example, a student may be required to achieve a 1 in order to enter the course. If he already can perform at level 6, he may not need the instruction at all.

The rating scale may also be used to inform a student of his progress. For example, he may be rated once a week throughout the course, and from these scores be able to pace himself accordingly. If students consistently fail to obtain the rating necessary to achieve the criterion performance, revision of the course curriculum may be indicated. Consistently low performance ratings require increasing amounts of revision. When a student achieves the criterion, no further instruction is necessary. Rating scales, however, require observers to score performance. So, the scoring is based on judgments, which sometimes makes the ratings unreliable. The more clearly specified the performance is at each rating scale point, the more reliable the ratings will be. Figure 3-7 shows a better rating scale for public speaking ability.

1	2	3	4	5	6
Is poor speaker but speaks without speech impediment	Has nervous mannerisms	Says "ah" a lot	Presents acceptable speech but delivery is too slow or is not sufficiently clear	Presents acceptable speech but is boring	Speaks fluently in a well-modulated voice, is interesting, does not pause inappropriately, etc.

Figure 3-7. Sample Behaviorally-Anchored Rating Scale

Nevertheless, errors are easily made on rating performances, so let's look at several different types of rating errors and ways to minimize them.

Since performance tests require the trainee to display actual outputs (product or process), they depend heavily on actual observations and rating of outputs. An examiner should rate performances or products under controlled conditions which should not change from one trainee to another. Also, the same performance standards should be used with each student. For example, a scale of 1 to 7 may be used to rate ability to drive a truck. Figure 3-8 shows such a scale with a rating of 4 specified as the standard acceptable for achieving the criterion.

1 2 3	4	5 6 7
	Rating needed to pass criterion test	

Figure 3-8. Sample Numerical Scale for Rating Driving a Truck

This standard should be the same for all students (A 7 means that the truck was driven in the best possible manner). A rating of 4 should mean that the truck was driven to minimum acceptable standards; ideally, all raters should agree as to what these standards are.

The problem of rating scales lies in the differing judgment of the observers. These differences (or rating errors) may be classified into four categories:

1. Error of Standards. Errors are sometimes made because of differences in observers' standards. If rating is done without any discrete, specified standards, there may be as many different standards as observers, thereby causing overrating or underrating. Standards at each point in the scale must be clearly specified. Consider the following example:

   > Ten persons are simultaneously being rated on their swimming ability. Judgments of the observers will, in this case, be dependent on their views of swimming standards and their relative experience in the area. The more knowledge and experience they have in the area, the more nearly alike their ratings of the students will be. More importantly, the more the swimming standards can be specified in terms of actual behaviors (for example, "legs do not bend at knees while kicking = 3"), the better the interrater agreement.

2. Error of Halo. An observer's ratings may be biased because he allows his general impression of an individual to influence his judgment. This results in a shift of the rating and is known as an "error of halo." If the observer is favorably impressed, the shift is toward the high end of the scale. If the impression is unfavorable, the shift is toward the low end. This type of error frequently goes undetected unless it is extreme. It is, therefore, a difficult error to overcome. Error of halo is reduced by reminding the rater that he is judging specific performances and should not take into consideration his impression of the individual as a whole.

3. Logical Error. A logical error may occur when simultaneously rating two or more traits. When an observer tends to give similar ratings to traits which aren't necessarily related, he is making a logical error. It may appear to him that these two traits are similar when they really aren't. It seems logical to him but more than likely doesn't to the other observers. For example, if "efficiency" and "productivity" are both being rated, some observers may think that they are highly related. Thus, they would tend to rate both traits at the same level: If a person is efficient, he must be productive. This isn't necessarily so, but a logical error is easily made in such cases.

The way to minimize logical errors is to make the distinctions among different traits to be rated as clear as possible. Point out to the raters that only separate, independent traits are to be rated. If possible, give examples of the behaviors associated with each trait.

4. Error of Central Tendency. An error of central tendency is demonstrated when different raters tend to rate most students toward the middle of the distribution. If, for example, the scale has seven points and you get a large number of 4s from your raters, they may be exhibiting an error of central tendency.

   One way to counter this is to use rating scales with an even (4, 6 or 8) number of points. Such scales have no midpoint and you therefore force raters to spread their ratings more than with a scale having a midpoint. The best solution, however, is to anchor your rating points with words which describe the behaviors and/or performances required (as shown in Figure 3-7).

Let's now look at a few specific types of process rating methods. There are several types of scales for rating performances that are observable but transient. You can use:

- A numerical scale
- A descriptive scale
- A behaviorally-anchored numerical scale
- A checklist

If at all possible, use the checklist. The checklist is generally derived from job performances and is the most reliable rating scale.

1. Checklist. A checklist is useful for rating ability to perform a set procedure. It's also a simple method of rating skills when your purpose is to see if students have reached a certain minimum level. The performance is broken down into elements, which allows the observer to indicate whether each step has been successfully achieved rather than merely whether or not final performance has been achieved. This helps to reduce the error of standards because it tends to minimize subjectivity. Instead of a large number of categories from which the observers may choose, there are only two, "go" and "no-go" on many different items.

2. Numerical Scale. A numerical scale divides performance into a fixed number of points (greater than two), depending on the number of discriminations required and the ability of the raters to make these discriminations. In most cases, observers can make at least five discriminations reliably, but not more than nine, so most numerical rating scales should contain five to nine points.

3. Descriptive Scale. The descriptive scale uses phrases to indicate levels of ability rather than numbers. Here, the discriminations can be varied to suit the performance, making such a scale more versatile than a numerical scale. However, there are also disadvantages. One major disadvantage is the interpretation of the phrases. A phrase may not mean the same thing to all observers. The more behaviorally descriptive the phrase, the better. Another disadvantage is the difficulty in selecting phrases which describe degrees of performance which are "equally spaced." For example, many observers consider "poor" and "fair" to be more closely related than "fair" and "good."

4. Behaviorally-Anchored Numerical Scale. The behaviorally-anchored numerical scale includes a numerical scale along with behaviorally descriptive phrases below each number. Both the number and the phrase must be considered by the observer. The description can be a single word or can be relatively detailed. The more detailed the descriptions, and the more they describe actual behaviors, the better the rating results are.

Types of Items for Performance Tests: Product Rating

Product rating is more reliable than process rating since a product is usually tangible. After completing a performance test, the product produced is compared with the required product. From this comparison, the rating is produced. This procedure minimizes many rating errors, since it provides the observer with a tangible standard with which to compare the product's suitability.

Product rating methods include the same main types as process rating methods:

- Checklists (go - no-go items)
- Numerical scales
- Descriptive scales
- Behaviorally-anchored numerical scales

For example, a product checklist for attaching a bayonet to a rifle might consist of items such as the following:

Circle one

- Is the bayonet firmly attached to the rifle? (go - no-go)
- Is the bayonet positioned properly? (go - no-go)

A behaviorally-anchored numerical scale for a product (correctly-gapped sparkplug) might look like this:

1	2	3	4	5
Sparkplug gap off by ± .004" of specified tolerance	Sparkplug gap off by ± .003" of specified tolerance	Sparkplug gap off by ± .002" of specified tolerance	Sparkplug gap off by ± .001" of specified tolerance	Sparkplug gap set at exact tolerance specified

Figure 3-9. Sample Behaviorally-Anchored Rating Scale

Example of Determining Item Format and Test Fidelity

Now that you are familiar with different types of items, and their advantages and disadvantages, you should be able to make a considered judgment of the type required for each of your objectives. When you decide what type of items your CRT should include and the necessary level of fidelity, document your decision so you can refer to it when you actually start constructing your CRT.

Let's look at an example of determining appropriate item format and test fidelity.

Assume that you are planning a CRT to cover a block of instruction on presenting oral briefings in a leadership course. The specific objective is:

- Given four hours of library research, be able to prepare and deliver a 10-minute briefing to a General Officer on the status of oil shale deposits as a major potential source of energy for the U.S. Army. The briefing must present clearly and succinctly the following topics:

1. How oil shale is formed
2. Where oil shale deposits are found
3. Potential products and uses of oil shale
4. Estimated amount of oil shale in the continental U.S.

Now, what test format do you use? You obviously do not have a spare General Officer available to practice on. An appropriate CRT format here might be an oral presentation to the course instructor scored on a go - no-go using a checklist to reflect appropriate aspects of coverage and presentation (a fairly high level of test fidelity). A test having a much lower level of fidelity (and certainly not recommended here) would be a paper and pencil multiple choice test on knowledge about oil shale deposits, and principles of oral presentation.

ITEM SAMPLING AND SAMPLING AMONG CONDITIONS

From Figure 3-1, you can see that once item format and level of fidelity are planned, the next consideration is whether or not items should be sampled for objectives. Item sampling within objectives should be considered when there are large numbers of items that could be created for an objective. If an objective calls only for a few specific items, (such as carrying out fixed procedures) there is no need to sample.

Sampling within objectives is often necessary in situations where the objectives to be tested involve abstract concepts. Examples of such abstract concepts include:

- Mathematical concepts (addition, multiplication, differentiation, vector analysis, etc.)
- Categorical concepts (identifying species of plantlife, recognizing symptoms of emotional disorder, selecting suitable positions for defensive fortification, etc.)
- Problem solving (be able to troubleshoot and identify the malfunction in any internal combustion engine)

Item sampling within an objective usually occurs in situations where the objective requires learning a concept (such as addition) as opposed to a process requiring a fixed order of doing things (such as folding an American Flag or issuing a call for fire).

In cases of teaching concepts it is generally not possible to develop test items for all possible examples of the concept. Consider the concept of addition. If the objective specified in the training program concerns learning to add two three-digit numbers, development of a series of CRT items which effectively tests all possible two-way combinations of three-digit numbers is virtually impossible. Hence, CRT items must sample from the population of items which could be generated to test the concept. We might, for example, develop five or six items, each of which call for the addition of two three-digit numbers, and assume that if the criterion had been met on these items, the student possesses adequate knowledge of the concept to generalize to any series of two three-digit numbers.

The more difficult it is to learn a concept, and the greater the number of possible items in the concept class, the more items will be required in your sample. In general, the more aspects there are to learn about a concept, the more difficult it is to learn.

Also, the more aspects a concept has that are similar to another different concept, the more difficult it is to learn. For example, if you are teaching people to recognize types of quartz, there are a number of aspects of quartz that you'll have to cover--hardness, shape, etc. There are also a number of aspects that quartz shares with other minerals--because of these similarities, teaching recognition of quartz will be more difficult: The student will have to learn to discriminate between quartz and other minerals having quartz-like aspects.

There are at least two other factors that affect the number of items necessary for sampling within objectives. First, the relative importance of a correct classification--whether or not the trainee has mastered the concept--should help determine the number of items necessary. If it is critical that a trainee master a concept, more items should be included for the objective to ensure that the trainee can accurately apply the concept. For example, in survival training, an individual must be able to distinguish between edible plants and poisonous varieties. So, a relatively large number of items requiring the individual to discriminate edible from nonedible plants is necessary.

Another factor that may affect the number of items required when sampling within objectives, is limitations imposed by practical constraints. That is, often time availability, costs, etc. may not allow you to include as many items as might otherwise be desirable.

Document your item sampling plan so you will have a record when creating items. This plan should describe the characteristics that the items to be sampled should have.

Should Performances be Tested Under Single or Under Multiple Conditions?

In many situations, CRT performances require testing under multiple conditions. You may need to perform certain tasks under both daylight and nighttime conditions, for example. As another example, astronauts must perform certain maintenance tasks both inside the spacecraft and during EVA (extra vehicular activity) outside the craft while tethered by a lifeline. You may have to perform tasks under overloaded conditions including high noise levels, humidity levels, temperature levels, and so forth.

One job which you as a CRT developer will have, is the determination of conditions under which your test will be administered. Your objectives will specify the condition or conditions required. Often, you may need to develop test items which could be administered under multiple conditions For situations in which performance must be exhibited under a large number of conditions, you may wish to devise a sampling plan to guide you in determining which conditions to develop test items for. (This assumes that it is impractical to test under all possible conditions.)

For each objective upon which a test item is to be constructed, you should examine the range of conditions stated. Next, you should make a list of these conditions and rank them in order of priority. Figure 3-10 presents guidelines for testing under multiple conditions.

When developing a scheme for sampling among a large number of testing conditions, rank the conditions in order of importance, and develop a CRT item for the performance under each condition ranked in the top 30 percent. The top 30 percent should include all the more critical conditions; if it doesn't, you may need to test under more than 30 percent of the conditions.

- If the performance must be exhibited under each of two conditions--you should develop a CRT item for each condition.

- If the objective states that the performance may be exhibited under either of two conditions, toss a coin and pick a condition.

- If the performance must be exhibited under three conditions--you should develop a CRT item which tests the performance under the two most important conditions.

- If the performance must be exhibited under a large number of conditions--you should develop a CRT to test the performance under at least 30 percent of the necessary conditions. Be sure to include the more critical conditions.

Figure 3-10. Multiple Testing Conditions

Let's consider an example: Assume an objective specifies testing marksmanship accuracy with an M-16. The trainee is allowed to fire 30 rounds of ammunition at a stationary target and must place at least 10 rounds within the bullseye. He must do this under the following conditions:

- Daytime and nighttime (illuminated range)

- Wind prevailing from left and from right

- Wind velocities of 0, 10 mph, 20 mph, and 30 mph.

These conditions combine sixteen ways, such as:

- Daytime, no wind

- Daytime, with 20 mph prevailing wind from the left

- Nighttime (illuminated range), with 30 mph prevailing wind from the right

- Etc.

Since there are a large number of conditions and you can't test under all of them (for practical reasons), you should develop CRT items to test marksmanship proficiency under at least 30 percent of them. Rank the conditions in order of importance, and develop CRT items for at least the top four items (30 percent of 16). Here wind velocity, direction, and day/night conditions are important. So, you may wish to develop items for:

- Daytime, with 30 mph prevailing wind from right to left
- Nighttime, at an illuminated test range, with 30 mph prevailing wind from left to right
- Daytime, no wind
- Nighttime, at an illuminated test range, with 20 mph prevailing wind from right to left

By testing under the more difficult conditions, you can usually be sure that the trainee can perform under the easier conditions. In this example, though, one easy condition is included: "Daytime, no wind." Inclusion of this condition is an aid to diagnosis. That is, if you had only the more difficult conditions and the trainee failed to perform to standards, you wouldn't know if the failure was due to the difficulty of the conditions or just an inability to perform the target shooting in general. Thus, the easy condition provides a check.

Document your condition sampling plan so you will have a record when you create test items. The sampling plan should indicate the conditions (or combinations of conditions) under which the trainees will be tested.

DETERMINING HOW MANY ITEMS TO INCLUDE IN YOUR TEST, AND DOCUMENTING YOUR TEST PLAN

One task remains: You must decide how many items your test should include. The answer to the question "How many items should I create?" depends upon the objective: The more complex the objective (the more subtasks it includes) the more items will be required to test it. This is true, but it does not provide enough guidance in decision-making for the item developer. Two other basic factors govern the number of items to be developed:

- The variety of conditions under which the objective must be tested.
- The objective's level of acceptable performance, specified as standards.

The first factor, variety of conditions, has been covered in the preceding section in terms of sampling among multiple conditions. There are, however, objectives which do not specify multiple conditions, yet which may logically be testable under many conditions. For example, if an objective requires a pilot to be able to land a light plane on the main east/west landing strip at Dulles Airport in Virginia, we might be able to test the objective with one item (that is, actually requiring the pilot to land his light plane on that runway). But, if the objective requires the pilot to land on any paved airstrip, we must require the pilot to make as many landings as we feel are necessary--on various airstrips, under various conditions. In doing this, we are considering the range of conditions specified in the objective when we determine the number of items in the test. Develop as many items as are needed to demonstrate that the trainees can perform under the required conditions, sampling the range of objects the trainee must work with, and the range of conditions under which he must work.

The second factor, level of acceptable performance specified as standards, must also be considered in determining the number of items to include. You must include enough items to ensure that the standards are met. For example, suppose an objective states:

- Given the appropriate sparkplug wrench, be able to remove a sparkplug from a 1970 six-cylinder staff car in one minute.

To meet the standard as stated in this objective, a trainee needs only to remove one sparkplug in one minute. Suppose the trainee removes the plug in 59.5 seconds but he is rushing frantically. He passes the item, but you aren't sure that it isn't a matter of luck--you're not certain that he could do it every time. In a case such as this, you might want to include two or three items. Each item must match the objective though. Thus, you might plan three items:

- . . .remove the #6 sparkplug in one minute.
- . . .remove the #5 sparkplug in one minute.
- . . .remove the #2 sparkplug in one minute.

You must plan these items before the test, and not vary them during testing. Actually, you are modifying the objective to state: ". . .remove three sparkplugs. . .in one minute or less per plug."

Consider this objective:

- Given your position as observer and the position of the enemy and description of his materiel, issue an appropriate call-for-fire according to the SOP.

If the trainee gave a correct call-for-fire but stumbled in saying it, you might be unsure whether he can meet the objective. Thus, you might write several items, each requiring that a different call-for-fire be issued. Several items would also allow for a wider range of stimulus conditions--your position, enemy position, and description of enemy materiel could all be varied. Again, you are modifying the objective to achieve a more accurate measure of the standard--this must be done before the test is given. It is never proper to add items during a test administration.

So. . .let's recap the general conditions for determining the number of items to sample the range of performances and conditions. We must create enough items to satisfy ourselves that, if passed, the trainee has met the standards. We must also be certain that each item matches the objectives even if there are many items for a given objective.

Do not get yourself into the conceptual dilemma of stating that "even if the student performs these four items I would not be convinced he has mastered the objective." If you find yourself in this situation--write more items. On the other hand, the test writer must guard against writing large numbers of items which test extremely rare performances under untenable and hard-to-imagine conditions. Simply make sure that all objectives are adequately sampled, and that all conditions and performances are covered--without being unreasonable and without writing large numbers of nitpicking items simply to watch the students squirm. It is important, however, that you sample all objectives, cover the necessary performances and conditions, and adequately cover the standards.

The reliability of your test--the extent to which it will measure the same thing each time you give it--is influenced by the length of the test. The more items you have on a test, the more data will be available for making determinations about test reliability. (Reliability is covered in detail in Chapter 7.) A good rule of thumb is: Write too many items rather than too few. You can use those which are left over to develop parallel, or alternate forms of the same test, or you can conduct an item analysis (as will be discussed in Chapter 5) and eliminate unnecessary and ambiguous items--keeping only the best ones for the final form of your CRT.

The Test Plan Worksheet

In developing a test plan, we have discussed:

- Overcoming practical constraints by selecting among objectives or modifying objectives
- Planning item format and level of fidelity
- Sampling items within objectives

- Sampling among multiple conditions
- Deciding how many items to include on the test

Figure 3-11 shows a worksheet which will help to collect and organize all the documentation of the test plan that you have developed. A worksheet such as this one should be developed for each objective upon which you wish to construct a CRT item. Figure 3-12 shows a sample worksheet filled in for the objective:

- "Given a set of climber's spikes and a safety strap, be able to climb a 30 ft. telephone pole in 2 minutes,"

as well as for two other related objectives.

Note that you should fill in the "number of items" column with the number of items required on the final version of your test. As you will see in the next chapter, you will create more items than this so that you can select the best ones by review and other techniques. By creating such a worksheet, you will have all the information needed for developing a test.

Objective (List Multiple conditions separately)	Select Among Objectives? (Indicate which ones or specify plan for selection)	Format	Fidelity Level	Type of Measurement (Process/ Product)	Type of Scoring	Sample Among Multiple Conditions? (Indicate which)	Sample Items Within Objective? (Show sample item)	Number of Items for Objective

Figure 3-11. Test Plan Worksheet

Objective (List multiple conditions separately	Select Among Objectives? (Indicate which ones or specify plan for selection)	Format	Fidelity Level	Type of Measurement (Process/ Product)	Type of Scoring	Sample Among Multiple Conditions? (Indicate which)	Sample Items Within Objective? (Show sample item)	Number of Items for Objective
Climb a 30 ft. telephone pole in 2 min.	No--all objectives are to be tested	Hands-on performance	High	Product	Go - No-go (Checklist)	No--only 1 condition	No--test item is only item. Item=climb pole in 2 min.	1
Correctly fasten climbing spikes to high top combat boots in 1 min.	No--all objectives are to be tested	Hands-on performance	High	Process and Product	Go - No-go (Checklist) on both process and product measure	No--only 1 condition	No--test item is only item. Item=fasten spikes to boots in 1 min.	1
Rotate 180° around telephone pole at height of 30 ft. Rotate to left. Rotate to right.	No--all objectives are to be tested	Hands-on performance	High	Product (The process is the product)	Go - No-go	No--test both rotating to left and rotating to right	No--test items (right & left rotation) are only items. Item=rotate 180° at top of pole	2 (rotate to left, rotate to right)

Figure 3-12. Sample Test Plan Worksheet

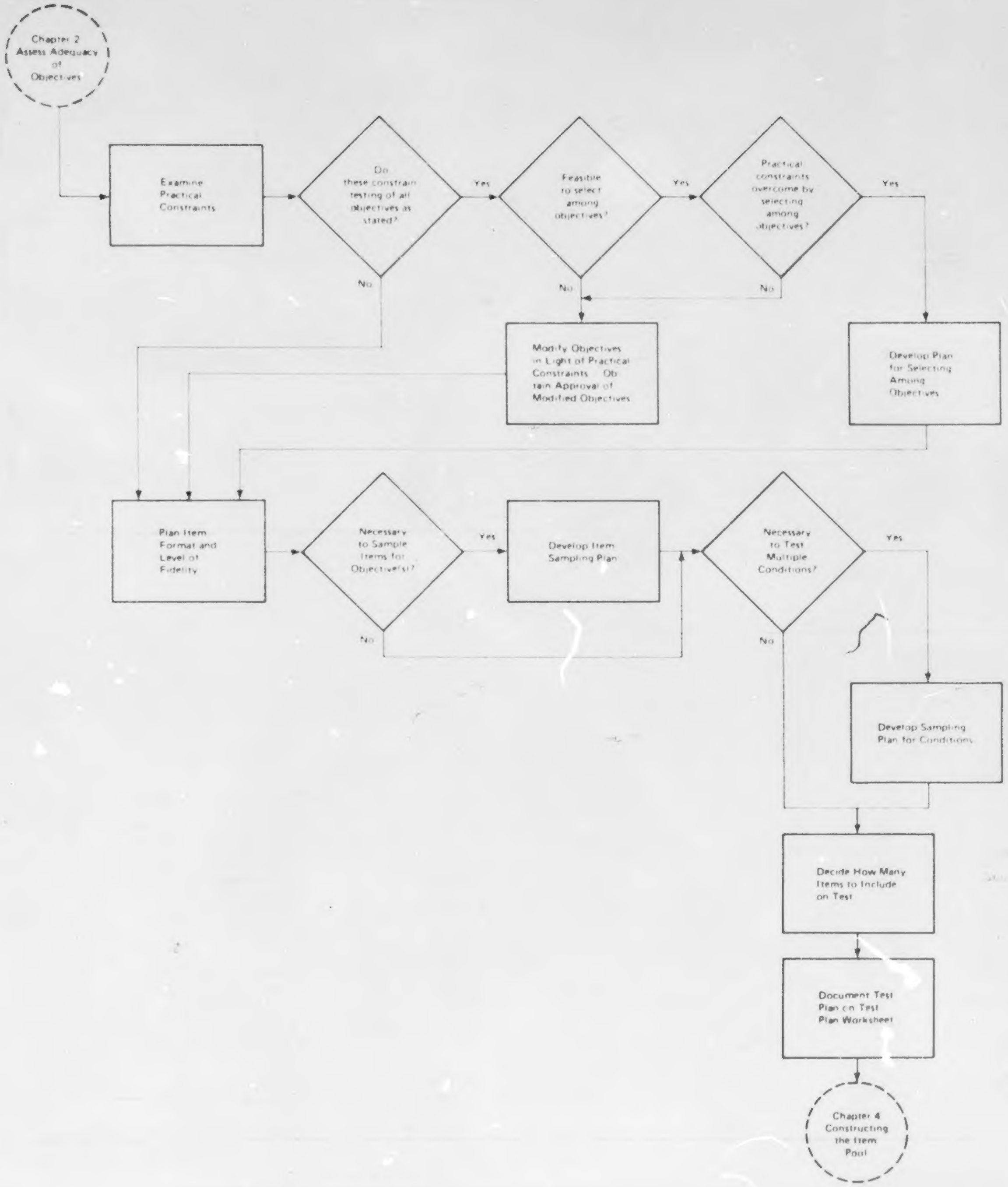

Chapter 2 Assess Adequacy of Objectives
Examine Practical Constraints
Do these constrain testing of all objectives as stated?
Yes
No
Feasible to select among objectives?
Yes
No
Practical constraints overcome by selecting among objectives?
Yes
No
Modify Objectives in Light of Practical Constraints. Obtain Approval of Modified Objectives
Develop Plan for Selecting Among Objectives
Plan Item Format and Level of Fidelity
Necessary to Sample Items for Objective(s)?
Yes
No
Develop Item Sampling Plan
Necessary to Test Multiple Conditions?
Yes
No
Develop Sampling Plan for Conditions
Decide How Many Items to Include on Test
Document Test Plan on Test Plan Worksheet
Chapter 4 Constructing the Item Pool

Figure 3-1. Sequence of Operations for Developing a Test Plan

CHAPTER 4

CONSTRUCTING THE ITEM POOL

"Constructing the item pool" is the process of creating a group of items from which final test items will be selected. The test plan, developed in the preceding chapter, documents the characteristics of the items necessary for your test. You have specified in your test plan the number of items required for each objective. You should create enough items to satisfy yourself that, if passed, the trainee has performed to the required standards under the appropriate conditions. It is advisable, however, to actually create about twice as many items as specified in the test plan. This will give you the latitude to choose the most appropriate items from a large item pool rather than to settle for the exact number you have created. You can tryout and review the item pool, and select among the items. In addition, extra items can be used to create alternate test forms.

Where the test plan calls for one item, you should build two; where it calls for two, you should create four. Thus, if the test plan specifies that an objective requires four items, two under each of two conditions, you would construct eight items--four under each of the two conditions.

Figure 4-1 (foldout at the end of this chapter) shows the sequence of operations necessary for constructing an item pool. Note that development of instructions is included as a part of this process: This applies both to instructions which tell the test administrator how to give the item (and test as a whole), and to instructions which tell the trainee how to take the item (and test as a whole).

CREATE ITEMS BASED ON TEST PLAN SPECIFICATIONS

The process of creating test items is relatively easy and straightforward, but calls for creativity and ingenuity. Take the test plan worksheet (completed in the operations described in the preceding chapter) and follow these steps to ensure construction of the appropriate items:

- Consider the first objective listed. If all objectives are to be tested, start with this objective. If there is a plan for selecting among objectives, start with the first objective specified for selection by the plan.

- Next, consider the format, fidelity level, type of measurement, and type of scoring specified for each item to be created for this objective. All items constructed for this objective must meet these specifications.

- Next, look at the worksheet column headed "Sample Items Within Objective?" This column indicates whether items will have to be sampled from a large group of appropriate items or not. If items must be sampled, this column indicates characteristics that each item requires.

- Then look at the "Sample Among Multiple Conditions" column. This column indicates the conditions under which each item must be tested. The column will specify how many conditions are to be tested and what these conditions are.

- Finally, look at the last column, "Number of items for objective " This column tells you how many items to create for each objective. Remember, if one item must be tested under two conditions, you create two items--one for each condition.

- Now, create the kind of items specified in your test plan worksheet for one objective. Then, repeat the entire process for the next objective specified on your test plan worksheet.

When creating items, first note the performance called for in the objective (overt main intent or indicator). Then write the test items following the test plan specifications, making sure that the performance in each item written for an objective matches the performance stated in the objective. You should be concerned not only with the performance (although the performance is extremely important), but also with conditions and standards. The rule for this is relatively simple:

> Make the test items include the same conditions and standards (no more, no less) as those specified in the objective.

Remember to consult your test plan, though--you may be sampling among the specified conditions.

Consider the following objective:

- Given a storeroom of tools used daily at the motor pool, identify the tools needed to replace a fanbelt on any late model jeep by taking those tools out of the storeroom.

Now suppose the item asks a student to remove tools from a dark storeroom at a specified motor pool. Would this be an adequate item? No! Who said anything about the storeroom being dark? The conditions called for in the test item are different from those called for in the objective. Not only the performance, but the conditions and standards also should be the same in the objective and the test item. That's the only way you will find out if the objective has been achieved.

When writing test items, remember to keep the language simple. The student's ability to comprehend difficult language is ordinarily not the skill in question. And remember, all indicators should be within the repertoire of the student. For example, if an item presents information to the student and requires him to calculate manpower needs for a tactical exercise, it should say "Calculate the required manpower" or "How many men are required?" Not, "Evaluate the logistical considerations and advance an estimate of personnel requirements pursuant to the information presented herein."

Now, let's consider an example of developing various types of items for the same objective. Assume that you have the following objective and must develop a test item:

> Objective: The student must indicate the best position for locating a light switch to activate a light in the supply closet of a batallion headquarters office.

One possibility is a standard multiple choice item. Figure 4-2 shows such an item.

The best place to locate a light switch for the supply closet is:

A. In the far left inside corner of the closet.

B. On the right inside wall of the closet about one foot from the closet door.

C. On the left inside wall of the closet about one foot from the closet door.

D. Outside the closet, about one foot from the closet door, and on the same wall as the door.

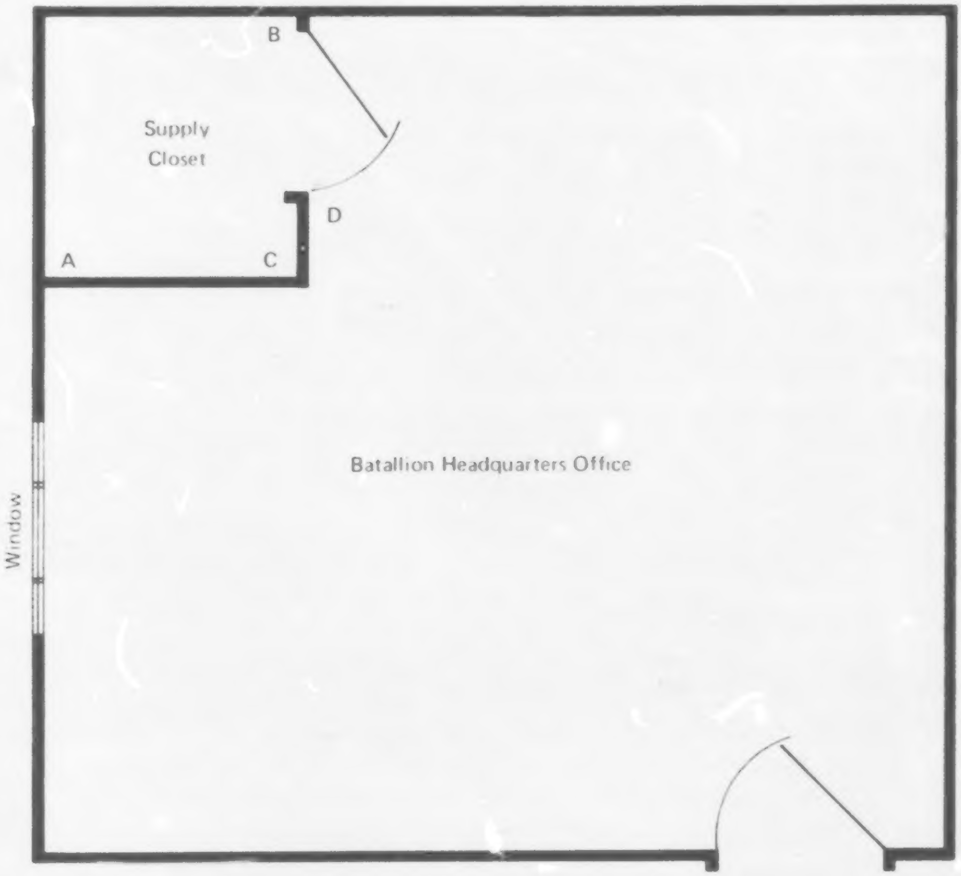
(Answer = D)

Figure 4-2. Sample Multiple Choice Test

However, this item requires the student to visualize the locations specified in the choices, A through D. The talent for this kind of visualization may not be in the students' normal repertories of behavior. This raises an important point:

> Use graphs, drawings, and photographs when necessary for clear communication.

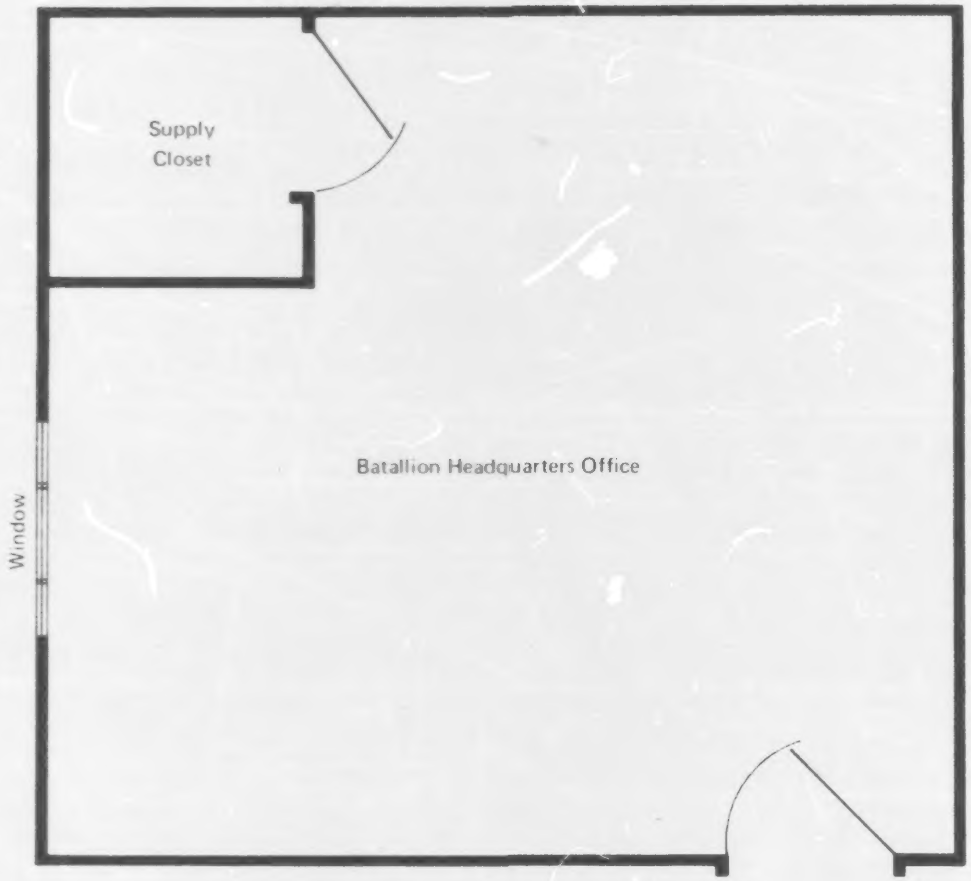
Keeping this point in mind, another, better possibility for the "light switch" objective is an illustrated multiple choice item such as that shown in Figure 4-3.



Directions: Place a circle around the letter which indicates the best position for the supply closet light switch.

Figure 4-3: Sample Illustrated Multiple-Choice Test

A third, even better, possibility is a simulated performance test item as shown in Figure 4-4.


Supply
Closet
Batallion Headquarters Office
Window

Directions: Place an "X" at the best position for locating the light switch to activate a light in the supply closet.

Figure 4-4: Sample Simulated Performance Test

Finally, the best choice is an actual performance item where the student enters the room with a red grease pencil and is instructed to "Place an 'X' at the best position for locating a light switch to activate a light in the supply closet." Of course, practical constraints may prohibit use of such an item.

Another point to keep in mind when creating items is the following:

> Present the test so it does not give the student hints as to the correct answer, but never make it extremely difficult simply to ensure a certain number of failures.

An example of a written item with a hint might be:

"An unfriendly force is shelling your position prior to attack. As soon as the shelling starts, your squad should begin a

1. Orderly retreat to get out of shelling range.
2. Attack to catch the enemy by surprise.
3. Advance toward the enemy position.
4. Move toward cover in previously prepared positions.

In this item, grammatical consistency gives a good hint. Choice 4 is the only one which grammatically follows from the item stem since "begin a move" is proper; while "begin a orderly," "begin a attack," and "begin a advance" are grammatically incorrect.

Remember your creativity and ingenuity are called for in creating items. You will have to use your imagination to create the best possible items for each objective.

DEVELOP AND DOCUMENT INSTRUCTIONS FOR ITEM USE

Once you have created the items for all objectives tested in your CRT, it is necessary to develop and document instructions that describe how each item is administered. Generally, tests consist of one type of item (performance items or multiple-choice items, for example), so instructions specific to each item are often not necessary--general instructions covering the entire test apply to all such items. (We will discuss general instructions in the last section of this chapter.)

Sometimes, though, specific instructions for each item are necessary. They may be necessary for two reasons:

- The item requires special equipment or facility setups, special conditions, or specific standards which the test administrator must implement as a part of administering that item.

- The item requires that special instructions be presented to the trainee in order for him to attempt it.

So, specific instructions are part of the items to which they are appended. The items could not be administered or understood without them. Thus, you must create specific instructions. Since they are a part of the item, item adequacy cannot be assessed without them.

When developing specific instructions, keep in mind the following points:

- Specific instructions should be placed with the items to which they apply. Those parts of the specific instructions which the trainee should read are written into the item. Those parts which tell the administrator what to do, should be included only in a separate "administrator's test copy."

- Specific instructions should tell the trainee whether speed or accuracy is more important. Any time limits should be specified.

- Provide clear instructions to the administrator. Tell him exactly what to say to the trainee, and how to answer questions. (The safest way is to have the examiner read to the trainee directly from the written directions.)

- Provide diagrams of equipment setups and facility arrangements for the administrator, whenever necesssary for a given item. Equipment settings (for example, dial settings on meters) should also be specified.

- Specific instructions should tell the trainee exactly what the performance, conditions, and standards are for the item--this is especially important for hands-on performance items. They may also explain the purposes of certain items. An example of a specific instruction is:

  "At this station you will be tested on your ability to perform certain tasks on the breech mechanism.

> These tasks will require you to perform the duties of several cannoneers. You have five minutes for each performance measure. You will respond appropriately when instructed. Using the breechblock holding tools and the eye bolts supplied, follow each instruction the examiner gives you. You must respond to each instruction correctly in order to pass the performance measure."

The administrator's specific instructions for this item would include what tools and eye bolts to assemble, how to place them at the station, and what instructions to give to the trainees.

Remember, an item is incomplete without necessary specific instructions.

After creating the items and their associated specific instructions, you should assess their adequacy. Let's review some of the requirements for adequate items.

ASSESSING ADEQUACY OF ITEMS

Do Items Match Objectives?

First, you should ensure that items match objectives. Check the following in both the item and the objective to be sure they are the same:

- Performances
- Standards
- Conditions

Then, find the overt main intent or indicator in the objective. This performance should be the same for each item you wrote for the objective. Do they match? If the answer is yes, proceed to the next check. If the answer is no, the item should be revised or rejected.

Third, note the standards in the objective. Make sure they match the standards in each item of the item pool for this objective. If they do not, the item should be revised or rejected.

Fourth, ensure that the conditions of the objective match those of the item. If they match, the item is successful. If they do not, the item should be revised or rejected.

Other Checks on Item Adequacy

You should also ensure that all items are clear and unambiguous. There should be no question as to what is meant. If you are not certain about any of your items or if you think that they can be taken more than one way, see if they can be improved by revision.

You should also take into account whether or not the items are reasonably easy to administer. An item should not be any more difficult to administer than is necessary while adequately matching the objective. Items that are complex to administer will be subject to additional error, both on the part of the test administrator and the trainees. For example, if your item is intended to assess beginning soldering skills, you would not want it to involve soldering microminiature components to a circuit board. Such an item would be difficult to administer because of the necessity of guarding against damaging expensive components, and because of the difficulty of observing the soldered connections. Instead, your item should probably involve soldering major components to a large chassis (or something similar which is more easily administered). The point is, not only should your items be feasible (be within the limits of practical constraints), they should also be relatively easy to administer.

You have stated in your test plan worksheet the level of fidelity as dictated by the test format. You should check now that your items are at the appropriate level. If your objective calls for hands-on performance, then your test plan worksheet should so specify. You must be sure that your items call for the same kind of hands-on performance.

Keep in mind that the higher the level of fidelity, the better the test. But remember, too, that the level of fidelity specified in the test plan must be adhered to, since it was based not only on the objective but also on practical constraints. (Practical constraints may have prevented higher levels of fidelity which would otherwise have been possible.)

When you revise inadequate items, be sure to revise their specific instructions also.

Now you have a pool of items and their associated specific instructions which appear adequate. All that remains is to develop general test instructions for your CRT.

DEVELOP GENERAL TEST INSTRUCTIONS

Proper instructions are an essential part of any test. You should try to make instructions as clear, unambiguous, and brief as possible--both general instructions given prior to the test, and specific instructions immediately preceding the items to which they apply. General instructions apply to the entire test, unlike specific instructions which apply only to certain items.

General instructions for any test should include the following types of information:

- The purpose of the test. For example,

  "This is a test of your ability to disassemble a M-60 machine gun";

  "This is a test of your ability to unscramble code words";

  "This is a test of your knowledge of traffic regulations";

  etc.

- Time limits for the test. For example,

  "You have 60 minutes to complete this test";

  "You have 40 minutes to complete Part A of this test, 30 minutes to complete Part B, and 45 minutes to complete Part C";

  "You should be able to complete this test in about one hour. Take your time, you will be allowed to finish if it takes you longer";

  etc.

- Description of test conditions. For example,

  "You will be allowed to use your textbooks";

  "You will be tested in a tent filled with CN tear gas";

  "You may use any of the tools on the table in front of you";

  etc.

- Description of test standards. For example,

  "You will be scored on how many items you complete correctly";

  "You will be scored on your ability to follow the SOP for doing this task";

  "You will be rated as to the smoothness of your landing";

  "To receive credit, you must get the exact answer for each problem";

  etc.

- Description of test items. For example,

  "For each problem, record your answer to the nearest tenth. Show your calculations";

  "Troubleshoot each malfunction listed and record the part to be replaced. Do one at a time, continuing until you have diagnosed each malfunction listed";

  "Circle the letter indicating the correct choice, A, B, C, D";

  etc.

  Note: If the test is a written one, it is a good idea to include a sample item with the correct answer. One sample item is worth many words of instructions.

- General test regulations. For example,

  "Do not talk to anyone--talking will cause you to fail the test";

  "Raise your hand if you need assistance";

  "Proceed to the next station when you have finished the task";

  etc.

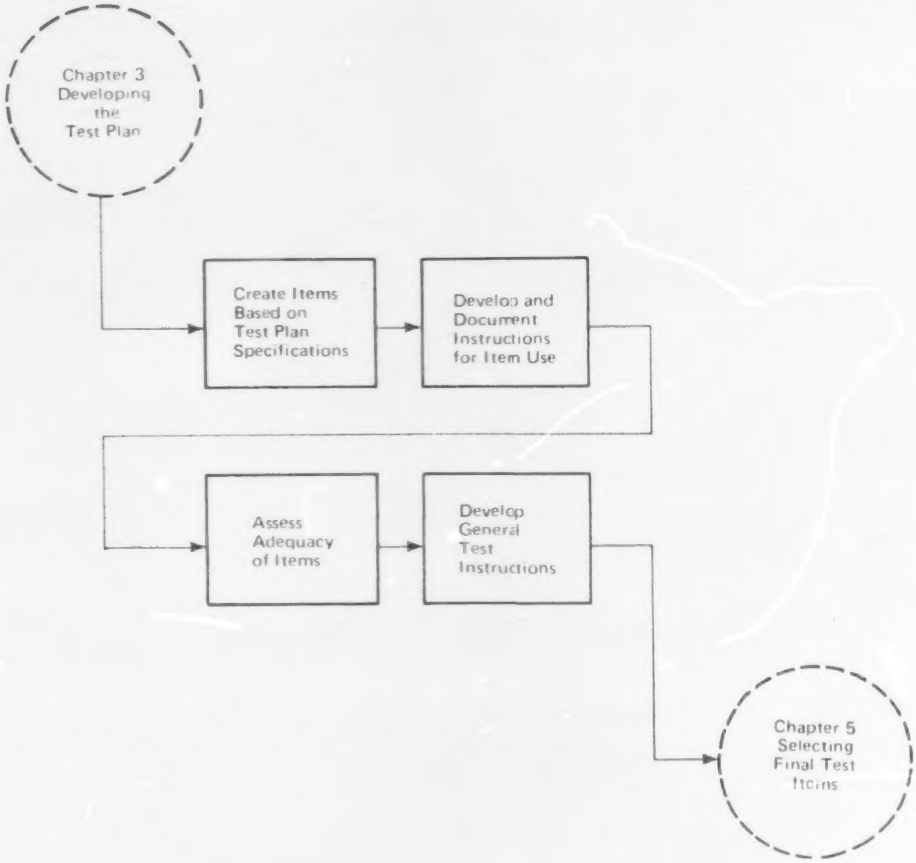

Chapter 3
Developing the Test Plan
Create Items Based on Test Plan Specifications
Develop and Document Instructions for Item Use
Assess Adequacy of Items
Develop General Test Instructions
Chapter 5
Selecting Final Test Items

Figure 4-1. Sequence of Operations for Constructing the Item Pool

CHAPTER 5

SELECTING FINAL TEST ITEMS

The preceding chapters have described how to construct test items for a CRT. The key characteristic of these items is that they are developed to measure the degree of attainment of an objective. The items you will select for the final version of your CRT will, therefore, depend primarily upon how effectively each item discriminates between those who have achieved the objective and those who have not. In addition, good items will not confuse trainees, and will pass reviews by peers and experts. The sequence of operations for selecting final test items from the item pool is shown in Figure 5-1 (foldout at the end of this chapter).

In order to select final test items, you will need a pool of about twice as many potential items as are required for the final version of your CRT. You have already checked each item to make sure that it matches its objective, and that the item is clear, unambiguous, reasonably easy to administer, and at the appropriate level of fidelity.

Even after such careful re-examination, it is important to try out the items. It is through tryout that problems which you cannot anticipate may become apparent. In this chapter, we will discuss how to conduct an item tryout and how to use the results. In addition, we will discuss other ways of reviewing test items, to help you select the best ones for the final version of a test. The end product will be a final version of a CRT which is ready to administer.

TRYING OUT THE ITEM POOL

Selecting A Sample

The sample of persons you use to try out the test items must include persons who are similar to those for whom the test is intended. Here we must keep in mind the purpose of the items--to differentiate between those who have the knowledges and skills to reach the objectives on which the items are based, and those who do not. So, about half of your sample should be composed of people who are "masters"--that is, people who have already passed the course segment that your item pool is testing, or those who are known to be competent in the subject matter area, such as instructors, or others who are already known to be qualified. The other half should be composed of people who are taking, or are likely to take, the instructional material for which you are developing a test, but who have not yet passed the course, unit, or lesson in question. The second half of your sample, then, should be composed of people who will be taking the CRT, but who are expected to be "non-masters" (since they have not yet had the appropriate training). Thus, about half of your sample may be expected to do well on the items in your item pool while the other half should not.

Suppose you had developed an item pool for individuals who have completed the individual tactical training component of BCT. Who would you try out your item pool on? Half of your sample should have people who have already been trained and tested on this component of BCT. The other half should be composed of individuals who are in BCT but who have not yet been trained in the individual tactical training component.

Suppose your test is intended for experienced intelligence specialists. Again, half your sample should be composed of such specialists but the other half should be composed of people who have been trained as intelligence specialists, and who are not yet experienced. It would be inappropriate to use people who have not received any training as intelligence specialists, since the purpose of the test is to identify those intelligence specialists who have had experience, from those intelligence specialists who have not.

> Try out your item pool on the same type of people as those who will take the final version of your test. Half the people in your tryout sample should be "masters," and the other half should be "non-masters."

If your test will be given to several different groups, you should try out the item pool on samples of "masters" and "non-masters" from each group.

Sample Size

The number of individuals to include in your tryout sample must be given careful consideration. Including too many is rarely a problem; the difficulty lies in determining the minimum number of people necessary for the tryout. There are two factors to consider in making this determination:

- The number of items in your item pool
- The size of the population for whom the test is intended

The number of items in your item pool is the most critical factor. You must have more people in your sample than items in your tryout pool. Otherwise you won't be able to use the tryout results properly.

In general, you should have at least 50 percent more people in your sample than items in your pool.

For example, if there are twelve items in your tryout pool, you will need a sample of at least 18 people (nine "masters" and nine "non-masters"). If possible, it is better to have an even larger tryout sample.

The greater the proportion of people in your sample to items in your pool, the more likely it is that your item analysis results will be reliable.

The second factor to consider in determining the size of the tryout sample is the size of the population for whom the test is intended. The principle here is:

The tryout sample size should be proportionally related to the size of the population for which the test is intended.

That is, the larger the size of the population for which the test is intended, the larger the tryout sample should be.

To be representative, a sample should have enough people to reflect the composition of the test population. There are no set rules for relating the sample size to the size of the test population, but Figure 5-2 provides some guidelines.

If your test will be administered to about this many people during one cycle:	Then the number of people in your tryout sample should be about:
20 or less	8 to 12
30	12 to 15
50	15 to 20
100	25 to 30
200	40 to 50
500	70 to 80
1,000 or more	80 to 110

Figure 5-2: Guidelines for Choosing Sample Size

If the population for whom the test is intended is small, the sample size can also be small and still be effective. So, for small populations, the sample size is more likely to be set by the number of items in the item pool. For example, if the population for a specific CRT in one administration will be about 20 people, you can see from Figure 5-2 that eight people will be enough for the sample (you would actually select four "masters" and four "non-masters"). But if your test will have about six items, then your item pool will have about 12 items. Thus your sample should have at least 18 individuals (number of items in pool plus fifty percent).

If the test population is large, the sample size will be determined more by the size of the population than by the number of items in the test. Remember that the number of items is the most critical factor. So, never use less than 50 percent more people than items even if the sample could be smaller based on the population size.

There is one other important point in selecting a sample that will be representative of the test population:

> The tryout sample must be random.

This means that the individuals chosen from among all available people of the appropriate type should be selected by chance. If you use a random sample, you will have the best representation of the test population.

It is very simple to construct a random sample. First, obtain two lists of the appropriate types of people ("masters" and "non-masters") available for the tryout. Write the names of the "masters" on separate slips of paper and place the slips in a helmet. Shuffle the slips thoroughly and, without looking, pull slips out of the helmet. When you have pulled out as many slips as needed for the "masters" half of the sample, keep these and throw the rest away. Then, make slips for the "non-masters" and repeat the process, ending up with the necessary number of "non-masters." You will then have a random sample of the appropriate number of "masters" and "non-masters."

Let's consider an example of determining a tryout sample. A very likely sample could be students who are about to start a training cycle. One group could be pretested (that is, tested before training) and called "non-masters." The second group, could be posttested (tested after training and called "masters."

Determination of Test Tryout Samples: Illustrative Problem

The test is to be five items in length. The course cycle has 50 people. Determine the number of people to include in the test tryout and the number of items. Assume you will use students in a current training cycle to develop the test for the next cycle.

Solution:

1. A five-item test requires 10 items for the tryout pool.
2. A 10-item pool requires a minimum of 15 people in the test sample.
3. Fifty people in the course cycles calls for 15 to 20 people in the tryout.
4. Randomly select a minimum sample of 16 people for the tryout, since the same number of people should be in each group.
5. Randomly divide the 16 into two groups of eight each.

6. Administer the 10-item pool to eight "non-masters" before training begins.

7. Administer the 10-item pool to eight "masters" after the training cycle is completed.

Conducting a Tryout

Now that you have selected a sample, you are ready to conduct a tryout of the item pool. The tryout should be administered in a standardized fashion, just as if you were giving the final version of the test. (See Chapter 6 for a detailed presentation of how to administer and score tests.) The item pool used in the tryout is likely to take twice as long for a student to complete as will the final version of the test, since it contains about twice as many items.

Here are some conditions you should establish during the tryout of the item pool:

- If possible, have someone else administer the item pool tryout, so you can be free to observe the process and note problems.
- Individuals in the sample should be informed that they are serving in a tryout to help develop a test. They should be asked to make notes of confusing or ambiguous items, and of anything they don't understand.
- Essentially the same instructions that will be used with the final version of the test should be used. It may not be possible to make these instructions exactly the same, since the test instructions may be modified based on feedback from the tryout. Certain test items may be eliminated by the tryout and subsequent review, so instructions associated with them will also be eliminated.
- The tryout is also used to evaluate the instructions: Lack of clarity, ambiguity, etc. should be noted by individuals in the tryout sample, and the instructions improved. (It is important to test for knowledge and skill in the areas covered, rather than for understanding of test directions!) Also, remember to give everyone in the sample the same instructions--this is important for standardization.
- Test conditions should be the same for the tryout as they will be in the final version of the test. Do not try to short-cut the specified conditions as this will affect your tryout results. For example, if items require the use of a 250 foot high jump tower for parachutist training, use that tower, not a 40 foot high jump platform. If a test item calls for outside administration, give it outdoors, not inside.

- Each item should be administered just as it will be in the test itself. This means, for example, that if it requires three test administrators to administer the final form of the test, you should also use three test administrators in the tryout.

- Test standards should be the same in the tryout as in the final version of the test. You must be careful to score the items for the people in the tryout exactly as you will for the final version of the test.

The tryout should be conducted exactly as if it were the final version of the test. Be sure to administer the tryout in exactly the same way that the test will be given.

Conducting An Item Analysis On The Tryout Results

There are a number of techniques that can be used to help spot bad items. All make use of the following principle:

> Acceptable items discriminate between "Masters" and "Non-Masters." Unacceptable items are incapable of making such a discrimination.

One simple and widely used item analysis technique makes use of a statistic called a Phi coefficient (ϕ, for short). The data required to use ϕ are:

- Which people who fail an item are "Masters" and which who fail it are "Non-Masters."

- Which people who pass an item are "Masters" and which who pass it are "Non-Masters."

If you have these four bits of data available, you can calculate the value of ϕ for each item.

Calculating ϕ

Let's look at an example of calculating ϕ. Suppose you have planned to have four items in your test. You have built an item pool consisting of eight items. You obtain a proper sample consisting of 12 individuals (12 = 50 percent more than the number of items, and the population for whom the test is intended is fairly small). Figure 5-3 shows the results of your tryout.

Recall that it was suggested earlier in this chapter, that approximately half of the people in your tryout sample should be "masters" (that is, people who have already completed the training segment that your CRT is being developed to test or experienced people who are acknowledged "masters" in the area tested). The other half should be people whom you would not expect to be "masters" (that is, people who are not necessarily knowledgeable in the subject matter being tested, or who have not had the appropriate training).

Trainee	"Master" or "Non-Master"	Item Number *								Number of Items Passed
		1	2	3	4	5	6	7	8	
T1	M	P	P	P	P	P	P	P	P	8
T2	M	P	P	P	P	P	F	F	P	6
T3	M	P	F	P	P	F	P	P	P	6
T4	M	P	P	F	P	P	F	P	F	5
T5	M	P	F	P	P	F	P	P	P	6
T6	M	F	P	P	P	P	F	F	F	4
T7	NM	P	P	F	P	P	F	P	P	6
T8	NM	F	P	P	F	F	F	P	P	4
T9	NM	P	F	P	F	F	F	F	F	2
T10	NM	F	F	F	P	P	F	P	F	3
T11	NM	P	F	F	P	F	P	F	F	3
T12	NM	F	F	P	F	F	F	F	F	1
Number Passed - Masters		5	4	5	6	4	3	4	4	35
Number Passed - Non-Masters		3	2	3	3	2	1	3	2	19
Total Number Passed		8	6	8	9	6	4	7	6	54

*P = pass the item; F = fail the item

Figure 5-3. Results of Item Tryout

Now, let's compute the ϕ coefficient for the items in Figure 5-3. Look at Item 4. For Item 4, we need:

1. The number of "masters" who gave the correct answer to Item 4.
2. The number of "masters" who gave a wrong answer to Item 4.
3. The number of "non-masters" who gave the correct answer to Item 4.
4. The number of "non-masters" who gave a wrong answer to Item 4.

Figure 5-4 is a matrix which helps organize data to simplify computation of ϕ. Let's put these data for Item 4 into the matrix in Figure 5-4.

	Item 4		
	Fail	Pass	
Masters	B 0	A 6	A+B 6
Non-Masters	D 3	C 3	C+D 6
Totals	B+D 3	A+C 9	12

Figure 5-4. Organization of Tryout Results For Computing ϕ for Item 4

In the upper right margin you write the total of A+B--the total number of "masters." The lower right margin (C+D) then is filled in to show the total number of people in the "non-master" group. The bottom left margin (B+D) shows how many people failed the item, while the bottom right margin's total (A+C) shows how many passed the item. The marginal totals (both the right margin and the bottom margin) must equal the total number of people in the tryout sample.

It is important to set up this matrix exactly as shown in Figure 5-4. The ϕ technique will not work correctly if you don't.

Figure 5-5 shows item/test matrices filled out for each item shown in the tryout results presented in Figure 5-3. Compare Figure 5-5 to Figure 5-3 to see how the matrices in Figure 5-5 were filled in.

Item #1

	Fail	Pass	
Masters	B 1	A 5	A+B 6
Non-Masters	D 3	C 3	C+D 6
	B+D 4	A+C 8	12

Item #2

	Fail	Pass	
Masters	B 2	A 4	A+B 6
Non-Masters	D 4	C 2	C+D 6
	B+D 6	A+C 6	12

Item #3

	Fail	Pass	
Masters	B 1	A 5	A+B 6
Non-Masters	D 3	C 3	C+D 6
	B+D 4	A+C 8	12

Item #4

	Fail	Pass	
Masters	B 0	A 6	A+B 6
Non-Masters	D 3	C 3	C+D 6
	B+D 3	A+C 9	12

Item #5

	Fail	Pass	
Masters	B 2	A 4	A+B 6
Non-Masters	D 4	C 2	C+D 6
	B+D 6	A+C 6	12

Item #6

	Fail	Pass	
Masters	B 3	A 3	A+B 6
Non-Masters	D 5	C 1	C+D 6
	B+D 8	A+C 4	9

Item #7

	Fail	Pass	
Masters	B 2	A 4	A+B 6
Non-Masters	D 3	C 3	C+D 6
	B+D 5	A+C 7	12

Item #8

	Fail	Pass	
Masters	B 2	A 4	A+B 6
Non-Masters	D 4	C 2	C+D 6
	B+D 6	A+C 6	12

Figure 5-5: Item/Test Matrices Filled In For The Tryout Results Shown In Figure 5-3

Now, you are ready to calculate the value of ϕ for each item. Figure 5-5 shows the formula for calculating ϕ.

$$\phi = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

That is: the numerator of ϕ equals the value of cell A multiplied by cell D minus the value of cell B multiplied by cell C. The denominator of ϕ is the square root of the marginal totals multiplied together. ϕ of course, is the numerator divided by the denominator.

Figure 5-6. Formula for ϕ

Now let's calculate ϕ for Item #1. Looking at Item #1 in Figure 5-5, you find the following values:

A = 5
B = 1
C = 3
D = 3
A+B = 6
A+C = 8
B+D = 4
C+D = 6

Substituting these values in the formula shown in Figure 5-6, you get:

$$\phi \text{ for Item \#1} = \frac{5x3-1x3}{\sqrt{(6)(6)(8)(4)}} = \frac{12}{\sqrt{1152}} = \frac{12}{34} = .35$$

Similarly,

$$\phi \text{ for Item \#4} = \frac{6x3-0x3}{\sqrt{(6)(6)(9)(3)}} = \frac{18}{\sqrt{972}} = \frac{18}{31} = .58$$

Note: Appendix D shows the square roots of all numbers from 1 to 1,000. You can use this table to help in your calculation of ϕ.

Using ϕ

The range of values of ϕ goes from -1.00 through zero to +1.00. The value of ϕ for a specific calculation may be anywhere in that range. Figure 5-7 shows the range of values of ϕ.

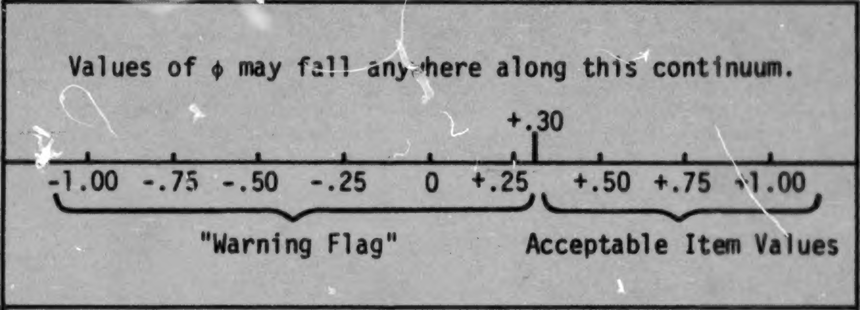

Values of ϕ may fall anywhere along this continuum.
+.30
-1.00
-.75
-.50
-.25
0
+.25
+.50
+.75
+1.00
"Warning Flag"
Acceptable Item Values

Figure 5-7. Range of Values of ϕ

The values of ϕ for all eight items in the tryout are shown in Figure 5-8.

Item #	ϕ
1	.35
2	.33
3	.35
4	.58
5	.33
6	.35
7	.17
8	.33

Figure 5-8. Values of ϕ for Items in Tryout Sample

If the value of ϕ is less than +.30 or is negative, the item may be a poor one. Regard values ranging from +.30 to -1.00 as "Warning Flags" that something may be wrong with the item.

A value less than +.30 means that the item does not discriminate very well between how masters and non-masters do. A negative value (-.55, for example) means that non-masters do better on the item than masters.

The values of ϕ for the eight items suggest that Item 4 is the best, followed by Items 1, 3, and 6 and then Items 2, 5, and 8. Item 7 in the example may be a poor item. Take a close look at this item before deciding to use it. (Your tryout sample may have been poor, or there may have been something wrong with the administration of the tryout, etc.) You should always regard an item with a ϕ coefficient ranging from -1.00 to +.30 with caution--something may be wrong with the item. A value of greater than +.30 indicates that the item is a candidate for inclusion in the test.

Summary of Using ϕ in Item Analysis

1. ϕ is best used when items are scored pass-fail, go - no-go, acceptable-unacceptable, or 1-0, and when there are about the same number of persons in the "Masters" and "Non-Masters" groups.

2. To compute ϕ for an item, determine:

   A. How many "Masters" passed the item.
   B. How many "Masters" failed the item.
   C. How many "Non-Masters" passed the item.
   D. How many "Non-Masters" failed the item.

3. Fill in the information determined above in a table such as this one (and make the additions indicated in the right and bottom margins of the table):

	Item		
	Fail	Pass	
"Masters"	B	A	A+B
"Non-Masters"	D	C	C+D
	B+D	A+C	

4. Calculate ϕ by substituting the values from the table into this formula:

$$\phi = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

5. If the value of ϕ for an item ranges from +.30 to -1.00, consider it a "Warning Flag" for that item: Pay careful attention to the item because it may be a poor one--it is often better to throw out that item, develop a new one and try it out.

Other Points About Item Analysis

ϕ may be used for conducting an item analysis of almost any CRT item pool. It is the technique of choice when the items are scored "pass-fail" or "go - no-go." However, ϕ can also be used when individual test items are given point values. In such cases it is necessary to set a "pass-fail" cut-off score for each item.

There are other related statistical measures which are more appropriate in other situations and scoring arrangements. These will be found in most standard books on elementary statistics.*

The ϕ technique described here is the recommended technique for computing item analyses. You should be aware, however, that if you have a very small sample, say less than 8 people (4 "Masters" and 4 "Non-Masters"), ϕ may not be appropriate. In such a case, you will have to resort to a more simple (and less accurate) technique.

Item Analysis by Inspection

If you have less than 8 observations, ϕ is inappropriate. In such a case, simply examine the numbers of "Masters" and "Non-Masters" who answered each item correctly. A rough interpretation about item selection can be made on the basis of judgments about these numbers relative to each other.

---

*For example: Guilford, J. P. Fundamental Statistics in Psychology and and Education. New York: McGraw-Hill, 1965.

Look at the data in Figure 5-3 for example. (Although we have more than 8 cases here, we can use this data to describe the procedure which is appropriate for small samples.) The best item seems to be number 4, with 6 "Masters" and 3 "Non-Masters" giving the correct answer. Items 1 and 3 look like the next best. Five out of 6 "Masters" passed these items, while 3 out of 6 "Non-Masters" gave the right answer. The fourth best items are 2, 5, or 8. These are marginal with only 4 out of the 6 "Masters" giving correct answers. Among these, the best choice would be that one which best rounds out the coverage of the selected items. Items 6 and 7 are the poorest of the lot. Only half of the "Masters" gave right answers to Item 6. It will need to be discarded or revised so more "Masters" will answer it correctly. There may be an unusual word or phrase in it which acts as a stumbling block. It may be necessary to create a new item to cover that objective. Item 7 shows too little discrimination between "Masters" and "Non-Masters."

You can see that these results correspond quite closely with the results of the ϕ calculations discussed earlier. Remember, the ϕ technique is preferred.

> You should only use the inspection method if you have less than 8 persons in your tryout.

Cautions on Use of Item Analysis Techniques

There are a number of cautions that you should bear in mind when using item analysis techniques on CRT item pool tryout results. These include the following:

1. An item analysis will only serve to warn you which items may be inappropriate for the final version of a test. It will not tell you which items are necessarily good. A low or negative ϕ does not mean that an item is definitely bad--it just means that you should consider it carefully before including it in your test.

2. Use the most appropriate item analysis technique that your data will permit. ϕ is the technique of choice unless your sample size is very small.

3. Some items may be "chained together" on certain tests. That is, they may all be a part of one performance measure. For example, a CRT on the disassembly of a specific weapon may have 10 steps, each of which is treated as an item and is scored go - no-go. Each of these steps must be completed in turn for the weapon to be adequately disassembled. But--if all steps are relatively difficult to perform (that is, some people fail

them, and some people pass them) except for steps 3 and 4 which are very easy, and which everyone passes, an item analysis would indicate that Items 3 and 4 have a very low value--probably around zero. That is, Items 3 and 4 in this case, do not discriminate well between "Masters" and "Non-Masters." Thus, you have a "Warning Flag" for each of these two items. But, you cannot throw out these items, since they are necessary steps in the disassembly of the weapon.

Whenever you have items that are "chained together" such as Items 3 and 4 in this example, you will not be able to throw some of the items out and keep others. You will either have to throw them all out or keep them all.

REVIEWING REMAINING TEST ITEMS

So far we have discussed only one way of selecting final test items: the use of item analysis techniques. Since item analysis will only provide "Warning Flags" concerning items which may be poor, you may require additional ways of judging items. Remember, since you have created an item pool of about twice as many items as your final test requires, your goal is to choose the best items for the final version of your test. It is not necessary to eliminate exactly half of the items in your pool, since you can always use extra items to make alternate forms of the test.

There are several ways in which you can review items in the item pool as supplements to the item analysis. They are all essentially subjective types of review and include:

- Feedback from individuals in the tryout sample
- Peer review
- Formal review by test evaluation units
- Formal review by subject matter experts

Feedback From Individuals in the Tryout Sample

Feedback from the individuals in your tryout sample can be extremely useful in helping you identify problem items. As discussed in the section on administering the tryout, students should write down misunderstandings, points of confusion, and ambiguities noticed during the tryout. You may want to use a worksheet, such as the one shown in Figure 5-9, to use in recording difficulties with the tryout.

Item #	Did you understand the instructions for this item?	Did you have enough time to do this item?	Did you understand how you would be scored on this item?	Were the equipment and facilities for this item suitable?
1 2 3 4 etc.				

Did you have any difficulties with the general test instructions? If so, what were they?

(Use as much space as necessary)

Describe any difficulties you had with items.

(Use as much space as necessary)

For each "no" in the table above, describe what the problem was.

(Use as much space as necessary)

Any other comments will be appreciated.

(Use as much space as necessary)

Figure 5-9: Worksheet for Recording Feedback From Tryout

If you use such a worksheet, point out to the individuals who complete it that their honest feedback will help you to improve the test. Note that the column headed "Did you have enough time to do this item?" is not relevant if you have items which involve time requirements or production rate standards. This column is intended to see if the individuals have enough time to complete items for which speed is not a part of the standard.

If many individuals (more than 20% of your sample) have difficulties with the same item(s), the item(s) in question may be poor.

If you have been able to get another person to actually administer the tryout for you so that you are free to observe, you should note the following points during the administration of the tryout:

- Did the trainees appear to follow the instructions easily? (If trainees appeared confused, you may want to ask them to repeat the instructions in their own words. If they can't do this adequately, make a note of the confusing instruction and revise it later.)
- Note questions asked by trainees. You may need to revise your instructions to take care of questions which come up frequently.
- Note problems with facilitites or equipment. Such problems may include malfunctioning equipment, equipment breakdowns, poor layout of facilities, hazards resulting from equipment or facilities, administrative difficulties in running trainees through the test on time, etc.
- If different performance measures are taken at different "test stations," note if there are any back-ups or bottlenecks going from station-to-station.
- Note whether the test administrator is able to adequately observe the performance of each individual. Also check to see if the administrator is inadvertently helping the trainees to do better than they could do by themselves.
- If you observe trainees making mistakes, talk with them to find out whether the mistake was due to a misunderstanding of the item or to an inability to perform.

You can use this record of observations to help discover poor items. In addition, some observations may aid in improving instructions, facilities, equipment, and other conditions of administration.

It is a good idea to have several administrators score each trainee independently. This is especially important if subjective rating scales are used. Note items which administrators consistently score differently--these may be poor items.

Peer Review

Another useful technique for evaluating items is to have peers review them. These should be fellow instructors, fellow test developers, etc. Ask your peers to review your item pool and to make notes of any items which they think should be revised or eliminated.

Formal Review by Test Evaluation Units

Another important type of item review is provided by test evaluation units. These units range from post educational advisors and their staffs to entire groups whose sole purpose is the evaluation of test materials. The test evaluation unit will be especially good at identifying problems with items that violate established testing principles. For example, they may easily identify items that are "give aways" or are too easy.

You should also give the test evaluation unit a list of the objectives, along with your item pool. They can then check to make sure that your items match your objectives.

Formal Review by Subject Matter Experts

Obtain a review of your item pool by subject matter experts. Since test evaluation units are often not experts on any particular subject matter (other than testing), you should obtain a separate review by subject matter experts for those tests on which you are not expert in the subject matter.

A subject matter expert can make sure that the content of your items is accurate. Request that the subject matter expert note any items which are confusing or misleading. Remember to give the subject matter experts your objectives, also.

REDUCING THE ITEM POOL

Now that you have completed an item analysis and submitted your item pool to a review, you are ready to reduce the item pool into a final test. Your goal here is to end up with a final test which incorporates the best items.

Fi[illegible]10 shows a simple way to summarize findings about items. In the "item analysis" column, check any items getting a ϕ from +.30 to -1.00. In the "tryout feedback" column, check the items with which a significant proportion of the people in your sample (more than 20%) had difficulty. Similarly, check the items which peers, test units, and subject matter experts agree are poor.

Item #	Item Analysis	Tryout Feedback	Peer Review	Test Unit Review	Subject Matter Expert Review
1					
2					
3					
4					
etc.					

Figure 5-10. Item Pool Review Summary Sheet
(Check items identified as poor)

Figure 5-11 shows a sample Item Pool Review Summary Sheet filled out for an item pool containing 10 items. Notice that Items 1, 3, and 4 appear to be okay: Neither the item analysis, nor feedback from the tryout, nor any other form of item review found fault with these items. Item 6 had a low ϕ value, but since no other form of review found fault with it, it is probably okay. Similarly, Item 7 may be okay, but you should check its structure--the test evaluation unit may have suggestions for approval. Item 9 was found poor by all techniques except tryout feedback; it should probably be eliminated.

Item 2 may have faulty structure since item analysis and the test unit review found fault with it, and since it confused the people in the tryout sample. Apparently its coverage of the subject matter was appropriate. Item 5, on the other hand, may have faulty content but acceptable structure.

Item 8 was found faulty only by the subject matter experts. Thus, it may have a technical error. Item 10, though, had a poor rating in the item analysis, caused confusion to the tryout sample, and was found faulty by the subject matter experts. This item should probably be eliminated.

In summary, Items 1, 3, 4, and 6 could be used in the final version of your test with no changes. Items 7 and 8 might be made acceptable with slight modifications, while items 2 and 5 would probably require greater efforts to make them acceptable. Items 9 and 10 should probably be eliminated.

Item #	Item Analysis	Tryout Feedback	Peer Review	Test Evaluation Unit Review	Subject Matter Expert Review
1					
2	✓	✓		✓	
3					
4					
5		✓	✓		✓
6	✓				
7				✓	
8					✓
9	✓		✓	✓	✓
10	✓	✓			✓

Figure 5-11: Item Pool Review Summary Sheet with Sample Entries for a 10-Item Pool (Check Items Identified as Poor)

The Item Pool Review Summary Sheet is just an aid to help you organize and consider the information you have collected about the adequacy of your item pool. Your own judgment must still play a major role, since you are more familiar with the items than anyone. So, using the Summary Sheet as an aid to your own judgment, you can decide which items are okay, which need improvement (and what kind of improvement), and which should be eliminated.

What To Do If You Eliminate Too Few Or Too Many Items

Often you may find that you have not been able to cut your item pool in half, or, on the other hand, that you have had to eliminate too many items. You don't really have a problem if you haven't been able to eliminate half the items in your item pool. In fact, you should be pleased--you have demonstrated your ability to create good items. What's more, you now have a choice: Either eliminate items by personal preference, or use the extra items to create alternate forms of your test. If you eliminate items by personal preference, be sure that you follow your test plan. For example, you may have planned a 12-item test with 4 objectives and 3 items per objective, and after reducing your item pool, find that you have 18 items with which to make the final version of your test. Be sure that you have 3 items per objective, after you discard the 6 extra items. Don't wind up with 6 items for 1 objective and 2 each for the other 3 objectives.

If you use the extra items to create alternate forms of your test, remember that alternate forms can share items in common. Suppose, for example, that you have eliminated only 2 items from an 8-item pool, and that the final version of your test requires only 4 items. Figure 5-12 shows the possible alternate forms of the test you can make with the 6 items, assuming that the items are independent and all are related to the same objective. Note that each of these fifteen forms has at least 1 item different from any other form. Each form, though, has at least half the items in common with any other form. Each form should be equally suitable as a final version of your test. (Note--there is no need for 50% overlap, it just works out that way in this example. If you had enough items left, you could create alternate test forms with no overlap. Such nonoverlapping versions are called "parallel test forms.")

If you eliminate too many items from your item pool, and don't have enough left for the final version of your test, you will have to create new items.

	Forms														
Item #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
2	✓	✓	✓	✓	✓	✓					✓	✓	✓	✓	
3	✓	✓	✓				✓	✓	✓		✓	✓	✓		✓
4	✓			✓	✓		✓	✓		✓	✓	✓		✓	✓
5		✓		✓		✓	✓		✓	✓	✓		✓	✓	✓
6			✓		✓	✓		✓	✓	✓		✓	✓	✓	✓

Figure 5-12. Alternate Test Forms Possible For a Four-Item Test Made From Six Items

If you must create new items, you should repeat the entire tryout item analysis and item review procedure using a new tryout sample and including the good items from the first tryout plus the new items. Often, though, you won't have enough time to do this. So, if you can't repeat a tryout using a new sample, try only the new items on your original sample. You can then compute new item analysis values for the new items. Then get feedback from the sample on the new items, and submit the new items for review by your peers, test evaluation unit, etc.

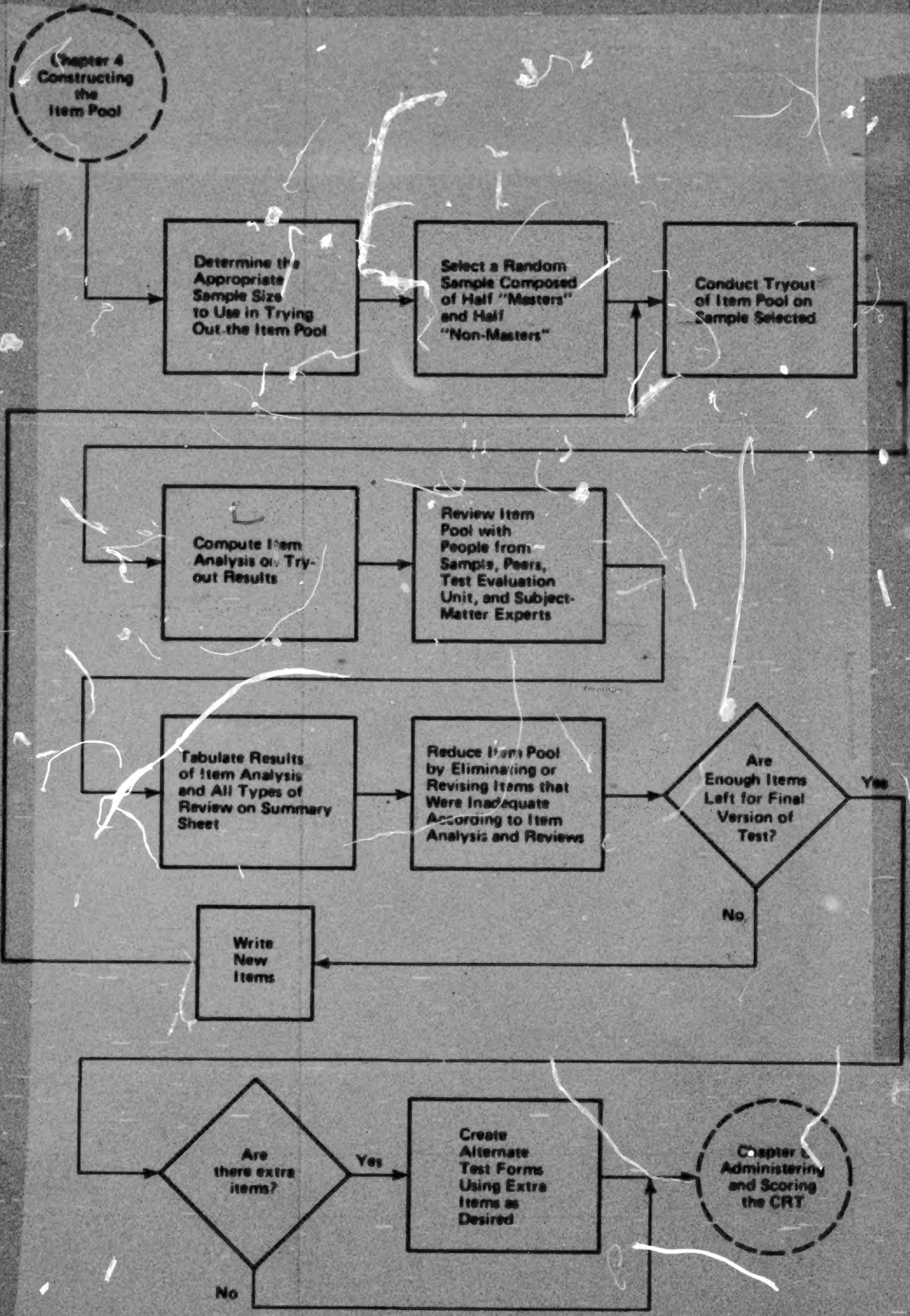

Chapter 4 Constructing the Item Pool
Determine the Appropriate Sample Size to Use in Trying Out the Item Pool
Select a Random Sample Composed of Half "Masters" and Half "Non-Masters"
Conduct Tryout of Item Pool on Sample Selected
Compute Item Analysis on Try-out Results
Review Item Pool with People from Sample, Peers, Test Evaluation Unit, and Subject-Matter Experts
Tabulate Results of Item Analysis and All Types of Review on Summary Sheet
Reduce Item Pool by Eliminating or Revising Items that Were Inadequate According to Item Analysis and Reviews
Are Enough Items Left for Final Version of Test?
Yes
No
Write New Items
Are there extra items?
Yes
No
Create Alternate Test Forms Using Extra Items as Desired
Chapter 6 Administering and Scoring the CRT

Figure 5-1. Sequence of Operations for Selecting Final Test Items

CHAPTER 6

ADMINISTERING AND SCORING CRTs

This chapter will familiarize you with procedures for administering and scoring CRTs. Efficient and objective methods of testing, accurate scoring, and fairness in interpretation of scores are essential in CR testing. This chapter will help you achieve these goals.

CONTROLLING THE TEST SITUATION

Although the use of a CRT implies that you are not interested in comparing the performance of one person with another, it is still necessary that interaction among trainees in the testing situation be prevented (unless, of course, the objective calls for the cooperation of two or more people). This simply means that, in paper-and-pencil testing for example, persons should be seated a reasonable distance from one another and within easy view of the supervisor; and that in group tests of performance, sufficient isolation should exist to ensure that students cannot help, hinder, or observe one another.

Whether testing is conducted individually or in groups, it is essential that test administration conditions be as nearly identical as possible on all testing occasions. This is necessary for proper assessment. For example, students should not differ greatly in their degree of fatigue, hunger, or on any other factor which could affect performance. The tester should also standardize his own behavior, his manner, and tone of speech when administering CRTs. Figure 6-1 (foldout at the end of this chapter) shows the sequence of operations for administering and scoring CRTs.

Controlling Environmental Variables

When administering CRTs, environmental conditions such as lighting, temperature, and background noise level, which might affect performance, should be standardized for all persons tested. For example, if the test involves visual acuity, the surrounding lighting must be very nearly the same from test-to-test. Conditions such as heat and humidity can seriously affect human performance, so that, especially for objectives requiring prolonged effort and concentration, groups tested at 72° F. might be expected to outperform equivalent groups tested at a humid 95° F.

Normally, the conditions required for testing should be stated in the directions. It is the responsibility of the tester to ensure that these conditions exist at the time of testing.

Controlling Personal Variables

Students should be tested under conditions comparable to those experienced by others who are tested. These include personal, physical, and emotional conditions. It would not be fair, for instance, to test one group of students for manual dexterity in the morning immediately following breakfast, and to give the same test to another group in the evening after a day of strenuous physical activity. Subjects complaining of minor illness may be excused and tested at a later time at the discretion of the test administrator.

Instructions and Tester Variables

Instructions must be uniform for all persons tested in order to minimize the possibility of cues and helpful hints becoming available to some persons and not to others. The standard test instructions should either be read, or recited from memory. Some typical and representative instructions for existing tests are shown in Figure 6-2.

The responsibility for standardization of test administration conditions rests with the test administrator. This includes standardization of your own behavior--the test administration procedures which you follow. For example, you are responsible for the proper timing and termination of the test.

In Chapter 2 the test designer was asked to keep in mind three main parts of a good objective: Performances, conditions, and standards. You, as test administrator, should also keep these components in mind. It is your responsibility to follow the specific guidelines for a given CRT.

Stated Test Objective	Instructions	Oral/Written Mode
1. Placing the M60 machinegun into operation and performing immediate action	"At this situation you must load the M60 and engage a target at ___ meters. You have three minutes."	Oral
2. Passage of obstacles at night and reaction to flares	"At this situation your unit is moving in the area of an enemy defensive position under simulated night conditions. You must cross a wire obstacle, a trench, and a danger area in order to reach your objective. Use nighttime techniques. Be prepared to react to an aerial flare."	Oral
3. Demonstrate an ability to comprehend written Russian by reading Russian prose passages and answering questions concerning them.	"In your test booklet you will find three passages from Russian novels. Read each passage carefully, then answer the multiple choice questions following them. You may go back and reread parts of a passage if necessary. You have 30 minutes to complete this test."	Written

Figure 6-2. Typical Test Instructions

Many objectives as written, are primarily product oriented. You should however, feel free to gather additional process information if such information appears to be useful in an auxiliary way, and can be obtained without interfering with the performance of those taking the test. For example, a trainee may be required to repair a radio/telephone. The "product" sought is an operational radio/telephone unit. "Process" information which might be noted includes style of work, care of tools, and adherence to approved procedures.

Figure 6-3 shows some typical steps which help ensure standardization of your own behavior in test administration procedures.

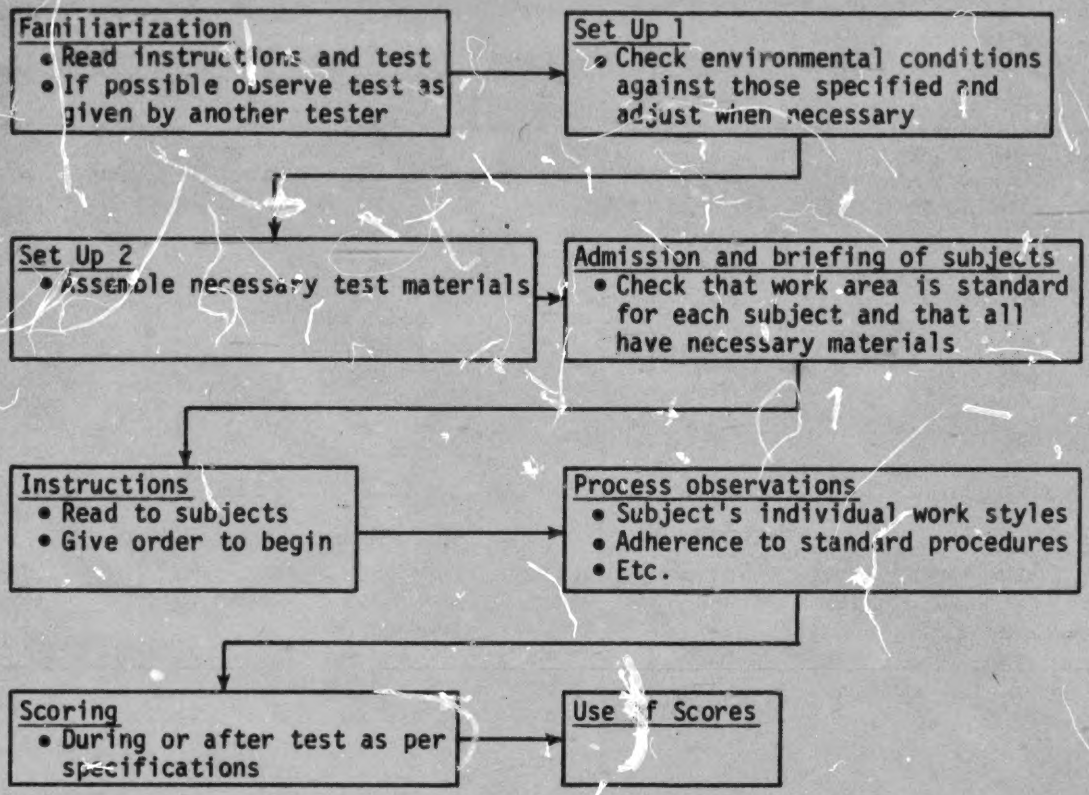

Familiarization
• Read instructions and test
• If possible observe test as given by another tester
Set Up 1
• Check environmental conditions against those specified and adjust when necessary
Set Up 2
• Assemble necessary test materials
Admission and briefing of subjects
• Check that work area is standard for each subject and that all have necessary materials
Instructions
• Read to subjects
• Give order to begin
Process observations
• Subject's individual work styles
• Adherence to standard procedures
• Etc.
Scoring
• During or after test as per specifications
Use of Scores

Figure 6-3. Some Typical Testing Steps

Remember, you must ensure standardization of all aspects of the test situation. Figure 6-4 summarizes the components of the test situation which you, as test administrator, must be sure are standardized.

Components	Examples
Environmental Variables	• Lighting conditions • Noise level • Temperature • Humidity
Personal Variables	• State of health • Time since rising • Time since last meal
Instructional & Tester's Variables	• Written or spoken instructions • Variations in tester work load (especially in group test situations when process observations must be made as well as product evaluations)

Figure 6-4. Three Components of the Test Situation

SCORING PROCEDURES

The aim of test scoring procedures is to obtain an accurate estimate of the trainee's competence. The less a test resembles a "hands-on" measurement the more difficult it is to reach an accurate performance measure. In cases where the measures are performance ratings, you should use several raters to judge the performance, rather than using a single observation. Be sure that raters are capable of making the judgments required. You are then in a position to assign scores with greater confidence, provided that the raters agree among themselves most of the time. If interrater agreement is very low, you should hesitate in interpreting the results. If interrater agreement cannot be achieved, the test items need to be reevaluated. (More about this in the "Rating Scales" section of this chapter.)

A number of different types of CRT scoring are currently in use. The proper scoring method is chosen with reference to a particular CRT, and with consideration of the complexity of the tasks and/or products required. The following sections discuss some common types of CRT scoring, including:

- Assist scoring
- Go - no-go scoring
- Fixed point systems
- Rating scales

Assist vs. Non-Interference Scoring

In CR testing, subjects generally proceed from the beginning to end of a test without comment or action on the part of the tester (non-interference). This type of scoring is often used in tests which call for the completion of a series of steps or which require production of a pre-specified product.

Some CRTs may, however, require scoring each step in a process. Thus, at each step, the student's performance is approved (scored "go") or he is assisted (and scored "no-go") before proceeding. Assist scoring may be employed for diagnostic reasons. Remedial training may then be focused on missed steps. This saves retraining time and expense. Assist scoring may also furnish valuable clues to areas where instruction might be improved. (A large number of errors in step number 3 of a 6 step procedure for example, may indicate an area where instruction could be improved.)

Example of Assist Method. After preliminary training, a food service course objective might require testing a trainee's ability to prepare a large meal. Here, it may be appropriate to observe each step in the cleaning, preparation and serving of the meal--correcting and recording errors as they are observed. If the entire sequence is carried out properly, the product measure will be scored "go." If errors are observed, the trainee may require additional training on the deficient steps. By using an assist method of scoring, not only is diagnostic information obtained, but a large meal is "saved"--the meal can be served. The trainee would be scored "no-go" if he was assisted on the test. However, the need for additional training before retesting would be minimized.

"Go - No-Go" Scoring

Generally, noninterference scoring is used with CRTs. The simplest noninterference scoring is "go - no-go" scoring. It is generally used to score simple, objective "hard-skill" processes or products. Since the score is either "go" or "no-go," the action must be performed (or the product assembled or created) exactly as specified by the objective. The item is essentially an observable expression of the standard in the objective. Either performance on the item meets the standard or it does not--there is no "gray" area.

Examples of Go - No-Go Scoring.

- A man is given 10 minutes to detect and replace a defective transistor in a radio set. He either does (go) or does not (no-go) have the unit operational within the allotted time.
- The assistant gunner on the M-102 Howitzer has the responsibility for setting the quadrant on the quadrant sight and firing the weapon. The required processes are:
    - Turning the counter handle to the appropriate numerical reading.
    - Raising or lowering the tube until the bubbles on the sight are level.
    - Firing the gun by pulling the lanyard on command.

Since this task can be precisely checked for accuracy, a passing score (go) is assigned only if no errors are observed on any of the above items.

Fixed Point Scoring

Another type of CRT scoring method is known as fixed point scoring. This type of scoring is appropriate when the task or product to be scored can be broken into several levels which may be quantitatively distinguished. For example, the item may call for adjusting valves to specified tolerances. If the trainee adjusts them to the exact tolerance, he gets 4 points. If he adjusts them to within ± .001 inch, he gets 3 points, ± .002 inch = 2 points, ± .003 = 1 point. No points are awarded if the trainee is off by ±.004 of an inch or more.

An alternate type of fixed point scoring uses "go - no-go" decisions on components of a task. For example, trainees may be asked to overhaul a carburetor, and a point value assigned to different components of the task:

Points	Task Description
1	Correct disassembly of carburetor
1	Correct cleaning of carburetor
1	Correct replacement of jets and parts of carburetor
1	Correct reinstallation of carburetor

A score of 4 indicates that all components of the task have been correctly performed. If the trainee failed to replace the jets and float but correctly performed components 1, 2, and 4, he would score 3 points on the task as a whole. A single test could test several tasks, each requiring performance on multiple components (subtasks).

Scoring is generally done using a checklist. All behaviors (or products) required by objectives are clearly defined. If the objective involves a product, scoring may compare the trainee's product with a sample product. For example, if an objective requires filling, sanding, and painting a dented metal surface to appropriate body shop standards, each finished product (the painted surface) is compared to standard products. The top standard is a smooth, high gloss metal surface. If the trainee's product is similar to this, he receives four points. The next standard is a smooth, high gloss metal surface with slight ripples. If the trainee's product resembles this, he gets 3 points. This progresses down to the zero point standard, which is represented by a metal surface which is finished so poorly that no points can be assigned.

Mixed Scoring Techniques

Sometimes several scoring procedures can be combined in one test. For example, suppose a test for the position of Radio/Telephone Operator has the following overall objective:

- "RTO (Radio/Telephone Operator) must be able to maintain the pack-mounted PRC-25 radio. Maintenance includes elementary troubleshooting, spot painting, periodic checks of rubber seals for cracks, and checking cable connections for fraying. The operator must demonstrate ability to translate and transmit frequencies and call signals of necessary units designated in the Signal Operating Instructions. He must also demonstrate ability to key the encoder with the Cryptographic Access Codes."

In this example, we can identify several objectives to be achieved by RTO candidates:

1. Ability to maintain equipment in working order
2. Ability to troubleshoot defective equipment
3. Ability to correctly identify incoming messages
4. Ability to accurately translate incoming messages
5. Ability to accurately encode own messages

So, we have broken down the duties of the RTO into 5 separate skill areas which may be tested and scored separately.

Objectives 1 and 2 might be scorable on a go - no-go basis. (Trainees are given a defective PRC-25 and uniform amounts of time to have their set operational.) Objectives 3, 4, and 5 however, might be scored on a point basis (go assigned for a score above a cut-off point but below 100 percent). If items pertaining to separate skills can be grouped and scored together, there is no real problem in testing an objective which is composed of different subtasks.

Rating Scales

Rating scales may be used to score CRTs, when dealing with more complex situations than those involved in "go - no-go" and fixed point systems. If the objective specifies characteristics of an acceptable action or product, a rating scale may be appropriate. Each item must be assigned a value on an explicit basis, so that independent raters will be able to agree consistently on their scoring. If possible, use two or more raters, who work independently.

To obtain a rough estimate of interrater agreement, line up the scores that each rater assigned each trainee on each item. Figure 6-5 shows an example for a six-item test taken by six trainees and scored by three raters using a 1-5 rating scale.

Looking across a row, you can compare the scores assigned by the different raters for each trainee. In the sample data presented, you can see that there is perfect agreement among raters on items one and five. On items two, three, and six, there is some disagreement. On item four, interrater agreement is very low--no raters agree on the score for any individual, and there is a range of four points between some ratings on that item. Thus, item four would either have to be drastically revised to increase interrater agreement, or dropped from the test.*

Item #	Trainee 1			Trainee 2			Trainee 3			Trainee 4			Trainee 5			Trainee 6		
	R_1^*	R_2^*	R_3^*	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3
1	5	5	5	3	3	3	4	4	4	2	2	2	5	5	5	1	1	1
2	5	4	4	4	4	4	3	4	3	1	2	2	4	5	5	2	3	2
3	5	4	5	4	3	3	3	3	3	3	2	2	4	4	4	1	1	2
4	3	5	2	3	1	4	2	4	3	1	2	4	4	2	5	2	3	1
5	4	4	4	4	4	4	3	3	3	2	2	2	4	4	4	2	2	2
6	4	4	3	3	2	3	4	3	4	3	2	2	3	3	3	2	1	2

*R_1=Rater 1, R_2=Rater 2, R_3=Rater 3

Figure 6-5. Comparison of Ratings on a 6-Item Test

The point system by which olympic divers are compared to an "ideal" dive (perfect performance of objective) is an example of a rating scale. Divers are not being compared directly to each other, but to a hypothetical "perfect performance" from which all divers fall short in some way or another.

In developing rating scales, the point assignment must be tied to criterion levels specified in the objective. If possible point assignments should be behaviorally-anchored. For example:

1 = does not complete job

2 = completes job in 45 minutes

3 = completes job in 30 minutes

4 = completes job in 15 minutes

5 = completes job in 5 minutes

---

*There are precise statistical techniques for measuring interrater agreement. For example, see:
Guilford, J. P. Psychometric Methods. 2nd edition. New York: McGraw-Hill, 1954, pp. 395-398.

Such behavioral anchoring will help to improve interrater agreement. The technique is, nevertheless, clearly more subjective than the fixed point system, and therefore, places additional responsibility on the tester. Ratings of ill-defined, global behaviors should be avoided entirely. For example, a rating scale with items such as "1 = does job poorly" and "5 = does job very well" would not be suitable since it would be likely to measure rater attitudes and opinions rather than the rated person's performance.

Figure 6-6 summarizes the three types of CRT scoring that we've discussed.

Type	Scoring Methods	Example
Go - No-go	Behavior performed correctly or not, product produced correctly or not	Trainee must jump trench after crouching and checking for sounds
Fixed Point Assignment	Points assigned to elements of a task with maximum score achieved when all items perfectly performed--maximum points assigned for a perfectly performed task or perfect product; no points are assigned if task is below minimum acceptable standards	In a complex first aid procedure such as wrapping a bandage, 1 point may be assigned for selection of the proper bandage, a second point assigned for wrapping the wound tightly, a third for covering the wound completely, etc.
Rating Scales	Numerical values attached by raters to a performance or product in which judgments of different raters may vary and therefore scores are not fully objective	Judging diving, or marching for form with values assigned to behavior on basis of its closeness to perfection

Figure 6-6. Types of CRT Scoring

Establishing Cut-Off Scores

CRTs are designed to assess proficiency on a given task or objective. Since it is often impractical to insist on complete mastery of the task (100 percent of items performed correctly) it may be necessary to decide upon a cut-off point (a score below which is considered failing or "no-go"). The more complex the skills assessed by the CRT and the more varied the type of performance or product, the greater is the danger of misclassification (designating a "non-master" as a "master," or vice versa).

There are no fixed rules or formulas for establishing cut-off points, but a number of factors can be considered:

- Immediate manpower needs--if manpower needs are very high it may be justifiable to lower cut-off levels especially if errors are less critical than no performance at all.
- Upper feasible score for an established "master"--a target may be placed so that even the best marksman may score only 50 percent hits. If we set a cut-off at 70 percent, we will pass no one at all.
- Criticality of the objective--the greater the risk of substantial damage to persons or to property, the higher the cut-off score should be.

> If a test is measuring more than one objective and cut-off scores are necessary, a cut-off level should be established for each objective.

For example, if one objective has four go - no-go items associated with it, the cut-off point for that objective might be passing any three out of the four items. Another objective in the same test may have eight items, with a cut-off score of passing any 6 out of the 8. Thus, a total of 12 points are possible on this two-objective test. If a person scores 9, he doesn't necessarily pass the test. He may have passed all four items associated with the first objective and failed 3 out of the 8 associated with the second.

Establishing cut-off points is a complex matter. You should reach a decision on this matter, only after careful consideration of the acceptable performance standards for the task(s) and task criticality. In general, cut-offs are useful when:

- Absolute mastery of the task is not expected but a suitable level of performance is specified in the objective.
- Absolute mastery is possible but factors other than competence affect the score (such as careless errors, measurement errors, etc.).

False Positives and False Negatives

The heart of CR testing is that "masters" must be correctly distinguished from "non-masters" in terms of specified criteria. It is important that competent people are not failed and that incompetent ones are not passed. Figure 6-7 outlines the concepts of "false positive" and "false negative" and shows possible results of such misclassifications.

Term	Definition	Possible Reasons for Error	Possible Consequences
False Positive	A trainee is given a "go" or point score above the cut-off but is really not a "master"	• Lucky guessing • Cheating • Selective preparation--test just "hit" the right items • Measurement error • Bias	• Damage to equipment • Personal injury • Inability to perform work properly
False Negative	A competent person who has in fact mastered the task is given a failing score	• Illness • Unknown behavioral fluctuations • Measurement error • Bias • Complexity of instructions	• Waste of training money • Possible unavailability of competent man because his skills are unrecognized

Figure 6-7. False Positives and False Negatives

Figure 6-7 shows that the consequences of either type of error may be extremely costly. Since CRTs may be employed to assess competence in widely varied tasks, it is difficult to make a general rule about appropriate places to set cut-off levels. However, a good guideline is specified below.

> If the cost of a false positive (passing an incompetent man) is very high, the cut-off point should be set very high.

This will eliminate trainees who are fairly competent (but not "masters").

One technique for reducing the numbers of false positives and false negatives, thereby reducing the likelihood of misclassification, is to increase the number of test items in use. It may be possible in some situations to increase the number of items simply by repeating the same item more than once (as in requiring student pilots to land a plane on a runway many times).

REPORTING AND RECORDING TEST RESULTS

Recording and reporting CRT results must be done in a precise, factual manner. After administering and scoring the test, the tester may, in addition, wish to obtain additional information. The following steps should be taken after dismissing the trainees from the testing situation.

- Retrieval and storage of relevant test materials, if any (pencils, answer sheets, rifles, dummy mines, etc.).
- Spot recheck of trainee's records for legibility.
- Recording of any additional process or product information which the tester observed and considers relevant to assessing the mastery of the task.

Behavioral observations which may shed light on the interpretation of test scores should be included with results whenever possible. For example, if trainees consistently complete all tasks on a go - no-go series in a very short time, this may be relevant to future training. On the other hand, a student may successfully get his radio in operational shape, but use an excessive amount of materials in doing so, or may damage the casing. Strictly adhering to the standardized scoring of the test might indicate a "go" score, but the tester may feel the task was carried out improperly. The correct course of action in this case is to score the individual according to standard procedures but to supplement the report with appropriate observations.

SPECIAL PROBLEMS

Standardizing format, administration conditions, and scoring of a CRT will minimize unusual problems. Nevertheless difficult cases may appear. For example:

- A soldier halfway through the only available form of a CRT develops an illness (or is for some other legitimate reason unable to continue). There is no second form of the test and the soldier has already seen the first form. What to do?

  or...

- CRT results for a group of men must be obtained immediately, but there is inadequate staff personnel to observe all of the process information required to assess whether objectives have been adequately met.

or...

- The CO requests the names of the 5 most skilled soldiers. The CRT shows 18 men with perfect scores. How are the honor graduates chosen?

Such problems are not internal to the CRT, but involve outside constraints or demands which cannot be met without weakening the standardization of the test or using it in a way for which it was not designed.

In situations such as these, you must decide, in conjunction with other interested persons, what are likely to be the costs and results. The man in the first example who developed an illness during the test might be observed individually in a "hands-on" situation to assess his competence. Or, when manpower needs are considered, this particular person may not be needed for that particular task. Answers to such questions can only be decided by personnel in a position to assess the needs of the program, the man, and the costs of various alternatives.

If special considerations seem to demand that testing is needed immediately (even if the standardization of scoring is below par due to a shortage of trained personnel, for example) the person requesting the immediate information should be informed of the dangers involved. If it is still necessary to administer the test under such circumstances, all scores are called into question, and this should be noted on the report. Ideally, a retest with an alternate form of the same CRT should be administered later.

Finally, as has been emphasized previously, it is not usually appropriate to use CRT results in a normative way (i.e., deciding who is best among those passing or worst among those failing). A NRT is called for in such cases. CRTs should be used in such a context only with the greatest caution, and preferably not at all.*

---

* See the section entitled "CRT or NRT" in Chapter 1.

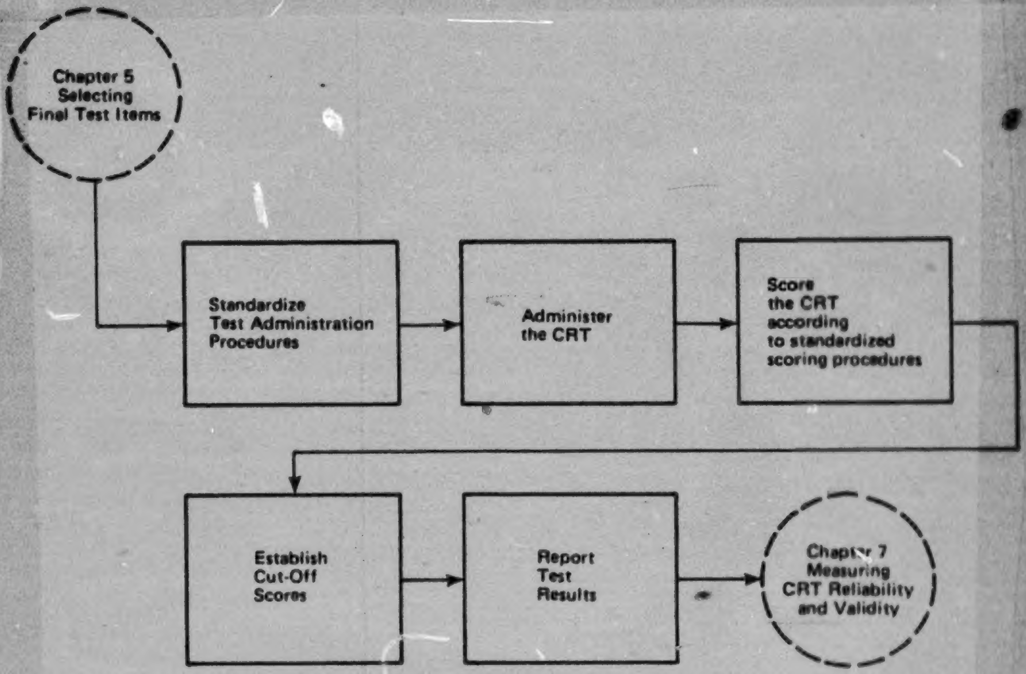

Chapter 5
Selecting
Final Test Items
Standardize
Test Administration
Procedures
Administer
the CRT
Score
the CRT
according
to standardized
scoring procedures
Establish
Cut-Off
Scores
Report
Test
Results
Chapter 7
Measuring
CRT Reliability
and Validity

Figure 6-1. Sequence of Operations for Administering and Scoring CRTs

CHAPTER 7

ASSESSING RELIABILITY AND VALIDITY

Two very important activities remain after you have developed your CRT--measuring the reliability of your test, and determining your test's validity.

Reliability refers to the extent to which a test yields consistent scores: If a test has high reliability, the same people should fail each time they take the test, while those who pass should do so consistently (assuming that no learning has intervened between test administrations). On a test which has low reliability, on the other hand, people of similar ability on the task may vary widely in their test scores, with some passing and some failing each time they take the test. If a test is highly unreliable, the same individual may pass it one day and fail it the next (or vice-versa) just by chance fluctuations. Thus, it is essential that your test be reliable: If it isn't, using it would be like using an altimeter which sometimes reads "+200 ft" when you're at 200 feet above sea level and sometimes gives the same reading when your are at 18 feet above sea level. The results of using an unreliable CRT are likely to be nearly as unfortunate as flying a plane with an unreliable altimeter and, conceivably, equally disastrous.

Validity refers to the extent to which a test actually measures what it is supposed to measure. For example, consider a multiple-choice paper-and pencil test on first aid procedures, developed as a low fidelity measure of ability to administer correct first aid treatment. This test may be reliable--that is, the same people may score about the same on it each time they take it (or take alternate forms of it)--but it is not necessarily valid. To determine if it is valid, you would have to determine whether a high score on the test means that a person can actually administer correct first aid treatment, while a low score means that he cannot. In other words, just because a test is reliable does not necessarily mean that it is valid.

On the other hand, a test which is not reliable cannot be valid. If a test does not give consistent results, it cannot be said to measure anything accurately. Consider the altimeter which sometimes registered "200 ft" at 200 ft above sea level and sometimes "200 ft" when actually at 18 ft above sea level. Is it a valid measure of height above sea level? No! It clearly is not accurately measuring altitude.

Suppose this same altimeter consistently registered "200 ft" when a plane was flying at 200 mph, "400 ft" at 400 mph, "50 ft" at 50 mph, etc. In a sense the altimeter is "reliable"--it gives the same results under the same conditions. But a wire is crossed somewhere, the altimeter is measuring airspeed--not what it is supposed to be measuring--altitude.

CRTs, of course, should be as reliable and as valid as possible. If you have followed the steps for the construction and administration of CRTs outlined in the preceding chapters, you have already gone a long ways toward maximizing reliability and validity. The steps presented helped you "build in" reliability by standardizing test conditions and by increasing the number of items in your test. The item pool tryout and review processes helped you increase reliability and validity by selecting the best and most consistent items. Matching the items to the objectives helped you maximize validity by assuring that the test items measure what they are supposed to measure.

Nevertheless, you cannot assume that your test is reliable and valid enough to be useful simply on the basis of having carefully followed the CRT construction process. There are many potential sources of error that can lower reliability and validity of the most carefully thought-out test. What you must do, is to determine your test's reliability and validity in actual use. This chapter presents techniques for doing that. Figure 7-1 (foldout at the end of this chapter) shows the sequence of operations involved in assessing reliability and validity.

ASSESSING RELIABILITY

The first thing to do in evaluating the usefulness of your test, is to assess its reliability. If it is not reliable, there is little sense in checking its validity. When you assess the reliability of a test, you are essentially asking "how consistent a measure is this test?"

A CRT, like any measurement device, has possibility for error in its use. Consider a ruler, probably the simplest type of measurement device: If you measure a person's height over 10 days, you would expect to get the same results on each day. But, there will always be some measurement error, even under the best, standardized conditions. So, the first day, you may find the height to be 5'9-5/32", the second day 5'9-1/8", the third day 5'9-3/16", etc. The extent to which your measurement is consistent over repeated trials defines its reliability.

Computing ϕ as an Estimate of Reliability

One good way to estimate the overall reliability of your test is to see the consistency with which people pass or fail it. The principle is:

> If the test is reliable, people who pass the first time should pass the second time, while people who fail the first time, should fail the second time.

Reliability estimates based on this principle are called estimates of test-retest reliability.

In Chapter 5, you saw how to compute ϕ for item analysis purposes. You can also use ϕ as a simple estimate of test-retest reliability. To do this, you should have a group of at least 30 people to whom you can administer the test twice. These people should be sampled randomly from the population of people who would ordinarily take this test. In order to estimate test-retest reliability properly, you need to test the same group of people twice, close together in time.

- You should let only about one day elapse between the first time you test them and the second time.

Another important point is:

> Do not tell the trainees that they will be tested again.

This is very important since you don't want students to practice between test administrations or try to recall the test in detail. Test-retest reliability assumes no practice between administrations and equivalent conditions both times. So, it is helpful if the trainees are kept occupied between administrations and don't have time to practice.

"Equivalent conditions" applies not only to the test environment but also to the trainees themselves--trainees should be equally rested, equally hungry, etc. during each administration. Thus, it is a good idea to test them at the same time both days.

To calculate ϕ for test-retest reliability estimates, set up your results from the two test administrations in a matrix such as that shown in Figure 7-2.

		1st Test Administration		
		Fail	Pass	
2nd Test Administration	Pass	B	A	A+B
	Fail	D	C	C+D
		A+D	A+C	

Figure 7-2. Matrix Used for Computing ϕ in Test-Retest Reliability Estimates

You fill out this matrix similarly to the way you filled out the item analysis matrices described in Chapter 5. In cell A, you enter the number of people who passed the test both times; in cell B, enter the number of people who failed the test the first time, but passed it the second time. In cell C, enter the number of people who passed the test the first time, but failed it the second time. And in cell D, enter the number of people who failed the test both times. The marginal total A+B shows the number of people who passed the second test administration, while C+D shows the number who failed the second time. B+D shows how many failed the first time, while A+C shows how many passed the first administration.

Figure 7-3 shows test-retest matrices filled out for two different tests. Let's use these matrices to calculate an estimate of test-retest reliability for each of the two tests.

Test A

(Administered to 30 people)

2nd Administration	1st Administration Fail	1st Administration Pass	
Pass	B 5	A 14	A+B 19
Fail	D 10	C 1	C+D 11
	B+D 15	A+C 15	30

Test B

(Administered to 40 people)

2nd Administration	1st Administration Fail	1st Administration Pass	
Pass	B 10	A 16	A+B 26
Fail	D 10	C 4	C+D 14
	B+D 20	A+C 20	40

Figure 7-3: Matrices for Test-Retest Reliability Estimates With Sample Data for Two Different Tests

Remember that the formula for computing ϕ is:

$$\phi = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Thus, for Test A,

$$\phi = \frac{(14)(10)-(5)(1)}{\sqrt{(11)(19)(15)(15)}} = \frac{135}{\sqrt{47,025}}$$

$$= \frac{135}{216.85} = .62$$

And, for Test B,

$$\phi = \frac{(16)(10)-(10)(4)}{\sqrt{(14)(26)(20)(20)}} = \frac{120}{\sqrt{145,600}}$$

$$= \frac{120}{381.58} = .31$$

So, Test A is more reliable than Test B, in terms of test-retest reliability. But, what value of ϕ indicates that a test is sufficiently reliable? A useful rule-of-thumb is:

> A ϕ less than +.50 indicates that the test is of questionable reliability. A ϕ of +.50 or more indicates that the test has sufficient reliability. (Remember that ϕ can range from -1.00 through 0 to +1.00).

Thus, test A in our example qualifies as reliable. Test B does not. Remember that +.50 is a rule-of-thumb and should not be followed rigidly. For example, if you found that one test had a test-retest reliability of .52, while another had a reliability of .48, you would not be justified in saying that the first was reliable and the second not.

ASSESSING VALIDITY

Once you have determined that your test has acceptable reliability, you can turn your attention to validity. A reliable test which doesn't measure the appropriate thing is no better than an unreliable test. There are three types of validity that are recommended for CRTs:

- Content Validity
- Concurrent Validity
- Predictive Validity

Each type of validity addresses the question "Does this test measure what it is supposed to measure?" in a different way. Figure 7-4 compares the three types of validity.

Type	How It Works	How To Determine
Content	Compares contents of test to objectives--Do items measure what the objectives say they should measure?	Systematically, but nonstatistically
Concurrent	Compares results on test to results on another measure of the objectives--Is success (failure) on test associated with success (failure) on another measure of the specified performance taken at the same time (concurrently)?	Statistically
Predictive	Compares results on test to results measured later on the job--Is success (failure) on test associated with success (failure) on another measure of the specified performance taken later, when the trainee is actually on the job?	Statistically

Figure 7-4. Three Types of Validity

Now let's discuss each of these types of validity separately.

Determining Content Validity

Content validity is probably the single best way of assessing whether or not your CRT measures what it is supposed to measure. In assessing the content validity of a CRT, you systematically check to see if each test item is measuring exactly what the associated objective says it should. If all items measure what the objective calls for, the test is content valid; if they don't, it isn't.* A simple example should help make this clear: Suppose you have a one-item CRT. The item and its objective are shown in Figure 7-5.

Objective	CRT (One-Item)
Given the appropriate tools, perform routine preventive maintenance on the 45 KW generator as specified in the operating and maintenance manual for same, within 30 minutes.	In front of you is a 45 KW generator and the appropriate tools. Perform routine preventive maintenance on the generator as specified in the operating and maintenance manual. You have 30 minutes to complete this task.

Figure 7-5. A One-Item CRT and Its Objective

Does this test have content validity? Well, performing routine preventive maintenance on a 45 KW generator (the test) is obviously the best measure of the objective (performing routine preventive maintenance on a 45 KW generator). So the test is content valid. That is, there is no better way to measure the objective than the test. Of course, if the objective itself was not properly developed, then the test is useless. That is, if the people you are testing are being trained to troubleshoot the generator, rather than to maintain it, the objective--and any test based on it--is inappropriate.

Content validity, then, is [illegible]tter of the extent to which a test corresponds with its objectives. Content validity is best viewed as absolute measurement. From an absolute point of view, the results of a CRT suggest that either an individual does possess the ability to adequately perform the task which the objective defines, or he doesn't. If the test items and objective(s) are precisely matched, the test is content valid. If all items are not precisely matched with their associated objectives, the test is not content valid. The items must be representative of all aspects of their associated objective. Thus, if the objective involves applying a concept which has three characteristics, the items must include all three characteristics.

---

*This assumes that the objectives themselves have been derived from an appropriate analysis of what the trainee must be able to do.

So, establishing content validity is simply a matter of systematically checking objectives and items. Basically, there are two steps involved:

- First, check to be sure the objectives have been properly derived from an analysis of what the trainees must know and/or do in order to perform the tasks for which they are being trained.
- Second, check each test item against its associated objective to see if the item measures exactly what the objective says should be measured. Be sure that the item covers all aspects of the objective.

If both checks are affirmative, your test is content valid.

If you have many items on your test associated with one objective, be sure that each item measures exactly what the objective indicates. If your test includes many objectives, each with more than one item, check each item against its associated objective. Do this systematically for each item, and you've assessed the content validity of your test.

You should be aware of the following principle:

> If objectives have been properly developed and the test consists of high fidelity items based on these objectives, your test will probably be content valid. If, however, the test consists of medium or low fidelity items, it probably will not be content valid.

So, if you have a high fidelity test, and a systematic check reveals that it does not have content validity, you are in trouble--something is wrong with the test. Either its objectives are not properly derived from a task analysis, or its items are not matched to the objectives, or both--back to the drawing board.

Whether or not your test has content validity, you should also compute statistical estimates of concurrent validity, predictive validity, or both. If your test is content valid, this further assessment will answer important additional questions, such as: "How does performance on the CRT compare to performance on another measure?"

If your test is composed of low or medium fidelity items and, consequently, has lower content validity, statistical estimates of validity are of primary importance. For example, suppose an objective states:

- "Be able to execute proper walking motions in a low gravity environment such as the moon."

and a one-item CRT developed for this objective states:

- "Make three steps in a gymnasium using the proper technique for a low gravity environment."

The item does not measure exactly what the objective calls for, so the test is not content valid. However, it may be valid in another sense; but to determine this, you will have to use either a concurrent or a predictive measure of validity.

Determining Concurrent Validity

Concurrent validity compares individuals' results on your CRT with their results on some other measure of the performance being tested by your CRT. Individuals take the CRT and the other measure close together in time (concurrently). The other measure must be the best available assessment of performance on the objective(s) in question. A statistical determination of the degree of association between results on the CRT and results on the other measure will provide an estimate of the concurrent validity possessed by the CRT.

Other measures commonly used to establish concurrent validity with a CRT include:

- Existing tests already in use
- Instructor ratings of students' performance
- Higher fidelity versions of the CRT being validated, and others

For example, a CRT on first aid techniques may be validated against instructor ratings of first aid achievement; or, it may be validated against an existing first aid test which has worked well. A multiple-choice CRT on vocabulary (such as: given a word to be defined, choose the best definition--A, B, C, or D) may be validated against a fill-in-the-blanks version of a vocabulary test (such as: here is the word to be defined, write a simple definition in the blanks below). The fill-in-the-blanks test is a higher fidelity measure than the multiple-choice test. Remember, though:

- The other measure must be a suitable one. If you don't have another measure which you consider suitable, you cannot establish the concurrent validity of your CRT.

Once you have chosen the other measure to use in establishing the concurrent validity of your CRT, the statistical determination is easy: ϕ is again appropriate.

Let's look at an example: Suppose you want to determine the concurrent validity of a new CRT on leadership skills. In the past, instructor's estimates of students' leadership skills have been used--reportedly with good results. To establish the concurrent validity of the CRT, have your sample evaluated for leadership skills by the instructor, then test them using the CRT. Record the results in a matrix showing the numbers of people passing and failing the CRT and the number of people rated acceptable (passing) and unacceptable (failing) by the instructor. Figure 7-6 shows such a matrix with sample data for this example.

Results of CRT

Results of Instructor's Ratings		Fail	Pass	
	Acceptable	B 6	A 36	A+B 42
	Unacceptable	D 16	C 2	C+D 18
		B+D 22	A+C 38	60

Figure 7-6. Matrix for Concurrent Validation With Sample Data

Then the ϕ for concurrent validity of your leadership skills CRT is:

$$\frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} = \frac{(36)(16)-(6)(2)}{\sqrt{(18)(42)(38)(22)}}$$

$$= \frac{576-12}{\sqrt{632,016}} = \frac{564}{795} = .71$$

You can use the same rule-of-thumb suggested for reliability estimated by ϕ:

If the ϕ estimate of concurrent validity is +.50 or higher, your CRT is probably of suitable validity. If ϕ is a value between +.50 and -1.00, your CRT is of questionable validity.

It is important to make sure that the following conditions hold when you establish the concurrent validity of your CRT:

- Your sample must be representative of the population for which the CRT is intended. (Again, random sampling from the population will accomplish this.)
- Your sample must be relatively large. A random sample of 50 to 100 people may be used, but you'll be better off using more than 100 people.

Determining Predictive Validity

Predictive validity is based on the same concept as concurrent validity, and can be estimated by ϕ in the same way. Unlike concurrent validity, though, predictive validity compares students' results on your CRT with their results on some other measure taken at a later time--when they are actually on the job for which they've been trained. Whereas the CRT and the other measure are taken close together in time for concurrent validity, they may be separated by six months or more for predictive validity.

So, predictive validity tells you the extent to which results on the CRT predict results on the job. Typical types of measures used in predictive validity (predicted by the CRT) include:

- Supervisor's ratings of on-the-job performance
- Other existing tests (such as MOS tests)
- Peer ratings of on-the-job performance
- Objective indices of on-the-job performance, such as amount of products turned out per day (acceptable or unacceptable), number of mistakes committed (acceptably few or unacceptably many), and others

You determine predictive validity using the same ϕ procedures as for concurrent validity. For example, you might validate students' performance on a CRT of leadership skills against supervisors' ratings of their leadership skills in their units six months later. Use the same rule-of-thumb as for reliability and concurrent validity:

> Acceptable predictive validity is defined by a ϕ greater than +.50.

The same cautions that apply to concurrent validity hold true for predictive validity:

- The measures against which you validate the CRT must be suitable--not just the only measures available. (If you don't have another measure which provides an acceptable assessment of on-the-job performance on the task tested by the CRT, you can't establish the predictive validity of the CRT.)
- Your validation sample must be representative of the population for which the test is intended.
- Your validation sample must be relatively large.

WHAT TO DO IF YOUR TEST RELIABILITY OR VALIDITY IS TOO LOW

As stated at the beginning of this chapter, your CRT must have both acceptable reliability and acceptable validity to be useful. In summary, here are the standards for judging the acceptability of your CRT's reliability and validity:

- Your CRT has acceptable reliability if the ϕ estimate of test-retest reliability is greater than +.50.
- Your CRT should be content valid, unless practical contraints have caused you to create a low fidelity test.
- Your CRT should have concurrent or predictive validity greater than +.50, as estimated by ϕ.

If your test does not meet these standards, it is probably not suitable for use as an Army CRT. Thus, you should either modify it or create a new test, and then assess reliability and validity again.

Following are some suggestions for modifying your CRT to increase its reliability and validity:

- You can often increase the reliability of a test by adding items. Of course, the items must match the objective(s). If the test is measuring several objectives, you must be sure to maintain the appropriate proportions of items to objectives. After you have developed and added items, reassess the test-retest reliability.
- A test that is not content valid due to lack of high fidelity items can be made content valid by reconstructing the items in a high fidelity format. You may have to modify practical constraints to do this, or make the test less feasible to administer conveniently.

But a difficult-to-administer, valid test is at least suitable for use, while an easy-to-administer test which lacks validity is unsuitable.

- If you have reason to believe that your test reliability or validity is too low because of improper sampling techniques, it may be appropriate to reassess the test using a new, more carefully selected sample. Be sure that the sample is properly large and representative of the population for which the test is intended. Also take care that the CRTs (and other measures) are administered in a proper, standardized fashion.

Do not misuse this last suggestion: Don't keep reassessing your test until you happen upon a time when reliability and validity check out as acceptable. You should only reassess if you think something was mishandled in the first assessment of reliability and validity, or if you modify the test. The test must be reassessed for reliability and validity after any and all modifications.

If you modify your test and it still doesn't have acceptable reliability and validity, it may be a good idea to seek help from your test evaluation unit. They may be able to see a difficulty that is not apparent to you--they may see the forest, when you've focused on the trees.

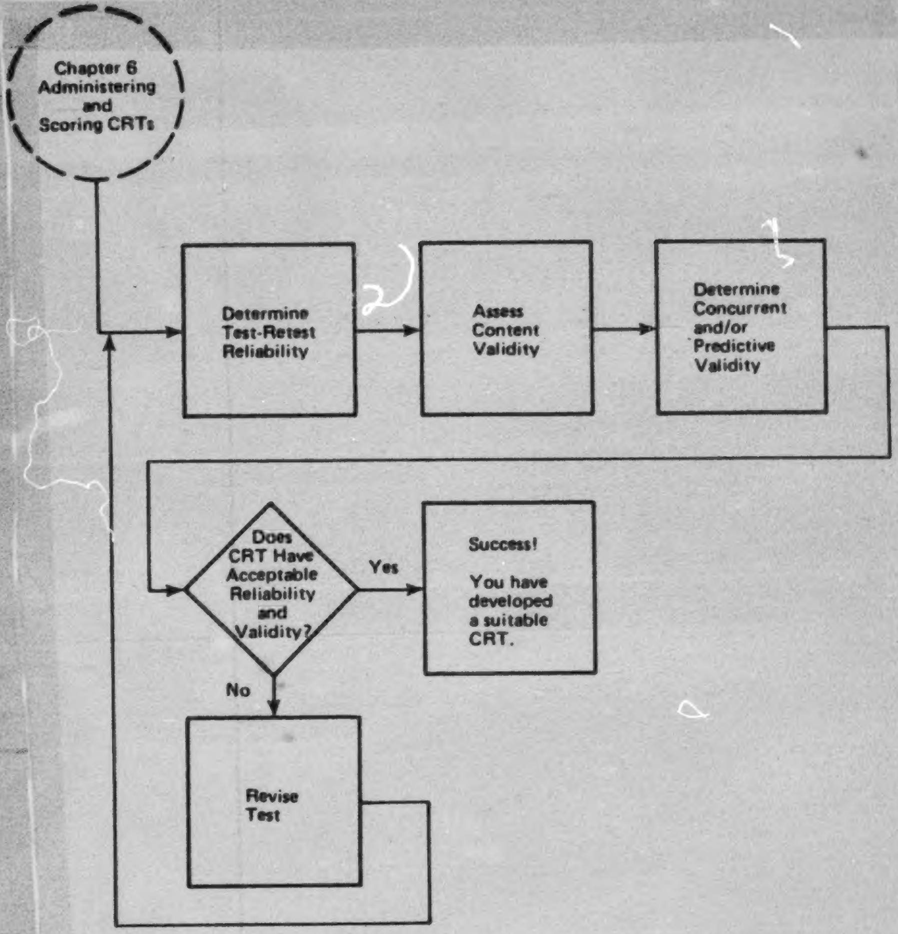

Chapter 6 Administering and Scoring CRTs
Determine Test-Retest Reliability
2)
Assess Content Validity
3
Determine Concurrent and/or Predictive Validity
Does CRT Have Acceptable Reliability and Validity?
Yes
Success! You have developed a suitable CRT.
No
Revise Test

Figure 7-1. Sequence of Operations Involved in Assessing Reliability and Validity

APPENDIX A

CHECKLIST FOR CONSTRUCTING CRTs

You can use this checklist to guide you through activities required to develop a CRT, once you are familiar with this manual. By using this checklist, you will be sure to perform all activities necessary for the development of an adequate CRT in the proper sequence. Consult the text if you require brushup information on activities. Remember, you should not use this checklist until you have gained familiarity with the CRT construction process by using the manual several times.

CHECKLIST FOR CONSTRUCTING CRTs

☐ 1. Determine whether a CRT is appropriate for required uses.

☐ 2. Determine whether a CRT can be built: Performance objectives external to training (what individuals should be able to do on the job) exist or can be specified.

☐ 3. Determine whether a CRT can be built: Test can be scored on an absolute basis—minimal standards for acceptable performance can be specified.

☐ 4. Obtain a list of objectives to be tested.

☐ 5. Check that objectives call for performance on just one task.

☐ 6. Check that all tasks are independent.

☐ 7. List the three main parts of each objective to be tested—performances, conditions, and standards.

☐ 8. Check that main intents of objectives are clear.

☐ 9. Check that performance indicators are simple, direct, and part of the trainees' repertoire of behavior.

☐ 10. Check that performances, conditions, and standards are specified in precise, operational terms.

☐ 11. Send inadequate objectives back through channels to their originator(s) for revision.

☐ 12. List practical constraints.

☐ 13. Assess practical constraints in terms of their impact on objectives.

☐ 14. Develop plan for selecting objectives, if appropriate.

☐ 15. Modify objectives, as necessary.

☐ 16. Send modified objectives through channels for approval.

☐ 17. Determine item format and level of fidelity.

☐ 18. Specify whether items will require product measures, process measures, or both.

☐ 19. Develop plan for item sampling, if appropriate.

☐ 20. Specify multiple conditions for testing.

☐ 21. Determine number of items to include in test.

☐ 22. Complete test plan worksheet, documenting test plan.

☐ 23. Write test items based on test plan specifications.

☐ 24. Develop and document instructions for item presentation and use.

☐ 25. Check to be sure that item pool includes about twice as many items as test plan specifies.

☐ 26. Check that items match objectives.

☐ 27. Check that items are clear, unambiguous, easy to administer, and at the proper level of fidelity.

☐ 28. Develop general test instructions.

☐ 29. Check that general instructions are clear, unambiguous and as brief as possible.

☐ 30. Select an appropriate sample for item pool tryout.

☐ 31. Check that item pool tryout sample is composed of "masters" and "non-masters."

☐ 32. Check that tryout sample size is at least 50% larger than the number of items.

☐ 33. Check that tryout sample is random.

☐ 34. Conduct item pool tryout.

☐ 35. Conduct an item analysis on tryout results.

☐ 36. Obtain feedback from individuals in the tryout sample.

☐ 37. Record comments from peer review of item pool.

☐ 38. Record comments from test evaluation unit's review of item pool.

☐ 39. Record comments from review of item pool by subject matter experts.

☐ 40. Summarize results from item analysis, tryout feedback, and various reviews of item pool on *Item Pool Review Summary Sheet.*

☐ 41. Reduce item pool, using *Item Pool Review Summary Sheet* as an aid.

☐ 42. Create and review new items if necessary.

☐ 43. Create alternate forms of test if appropriate.

☐ 44. Check that environmental, personal and tester variables are standard.

☐ 45. Administer the CRT.

☐ 46. Score the CRT.

☐ 47. Establish cut-off scores.

☐ 48. Report test results.

☐ 49. Collect test-retest reliability data on appropriate sample.

☐ 50. Calculate ϕ as an estimate of test-retest reliability.

☐ 51. Check that ϕ is greater than +.50.

☐ 52. Assess content validity of CRT.

☐ 53. Select an appropriate "other measure" for concurrent/predictive validation of CRT.

☐ 54. Obtain a relatively large, representative sample for use in evaluating concurrent and/or predictive validity.

☐ 55. Administer CRT and other measure to sample concurrently or, after appropriate interval, predictively.

☐ 56. Calculate ϕ as an estimate of concurrent and/or predictive validity.

☐ 57. Modify test to increase reliability and/or validity, if necessary. Following such modification, reassess reliability and validity of test.

APPENDIX B

CHECKLIST FOR EVALUATING CRTs

You should use this checklist to help you evaluate CRTs which have already been constructed. This checklist will help determine the suitability of CRTs which already exist, and which you may wish to adopt for your own testing needs.

This checklist consists of an ordered series of questions to ask when evaluating a CRT. Some of these questions pertain to physical aspects of CRTs and can be answered just by looking at the test, without knowing any additional information. Other questions concern CRT use: To answer these, you must know the objectives, intended test population, practical constraint data, reliability and validity estimates, etc. So, before using this checklist to evaluate a CRT, collect the documentation that was used to develop it.

Circle the "Y" beside a question if the answer is "yes." Also circle "Y" if the question is not applicable to the test. If the answer is "no," "can't tell," or "partly yes, partly no," circle the "N" next to the question. When you have completed the checklist, the circled "Ns" will represent a record of the particular aspects of the CRT that may require upgrading or that require further information before being evaluated.

CHECKLIST FOR EVALUATING CRTs

N Y 1. Are all three parts of the objective present?

N Y 2. Does each objective call for performance on just one task?

N Y 3. Are all tasks independent?

N Y 4. Are main intents of objectives clear?

N Y 5. If main intents are covert, do objectives include performance indicators?

N Y 6. Are performance indicators simple, direct and part of the trainees repertoire of behavior?

N Y 7. Are performances specified in precise, operational terms?

N Y 8. Do objectives include statements of conditions and standards specified in precise, operational terms?

N Y 9. Are objectives free from impact of serious practical constraints? That is, do objectives *not* require excessive time, manpower and costs, elaborate facilities/equipment, etc?

N Y 10. If selection among objectives has taken place, were the objectives chosen at random from the entire population of objectives available for testing?

N Y 11. Are the students who were tested unaware of the sample of items selected for testing?

N Y 12. Does the item format selected best approximate the behavior specified by the objective?

N Y 13. Is the measurement used the same as that which is required by the objective (product measurement, process measurement or both)?

N Y 14. Has the possibility of rating errors been held to a minimum?

N Y 15. Is the item format at the highest level of fidelity practicable?

N Y 16. If item sampling within objectives has taken place, has the appropriate number of items been included?

N Y 17. Is the performance being tested under all conditions or, if it is not possible to test under all conditions, under an adequate number of conditions (and the appropriate ones)?

N Y 18. Was there an adequate number of items included in the item pool to sample the range of performances and conditions?

N Y 19. Are conditions and standards stated in the objective reflected in the test or item?

N Y 20. Does the performance in the item match that in the objective?

N Y 21. Are all items clear and unambiguous?

N Y 22. Are items reasonably easy to administer?

N Y 23. Are items at the appropriate level of fidelity?

N Y 24. Has the language of the CRT items been kept simple?

N Y 25. Is the student informed as to whether speed or accuracy is more important?

N Y 26. Are graphs, drawings and photographs used when necessary for clear communication?

N Y 27. Is the test presented in a way which neither gives the student hints, nor makes it extremely difficult?

N Y 28. Are instructions common to all items included in the general overall test instructions?

N Y 29. Do general instructions for the test include the following information: purpose of the test, time limits for the test, description of test standards, description of test items, and general test regulations?

N Y 30. Do specific instructions tell the trainee exactly what the performance, conditions and standards are for the item?

N Y 31. Are clear instructions provided to the examiner?

N Y 32. Have the items been "tried out"?

N Y 33. Was an appropriate sample used in the tryout?

N Y 34. Was the tryout sample composed of "masters" and "non-masters"?

N Y 35. Was the sample size at least 50% larger than the number of items?

N Y 36. Was the tryout sample random?

N Y 37. Was a "proper administration" of the tryout conducted?

N Y 38. Was an appropriate item analysis used?

N Y 39. Were additional evaluation techniques used to supplement item analysis (including feedback from individuals in the tryout sample, peer review, review by test evaluation units or review by subject matter experts)?

N Y 40. After item analysis and review, were poor items deleted or improved and only the best items used?

N Y 41. Is standardization of environmental, personal and tester variables specified in the directions?

N Y 42. Was the proper scoring method chosen with reference to this particular CRT?

N Y 43. Are the scoring procedures clear?

N Y 44. Were appropriate cut-off scores established?

N Y 45. Was a cut-off level established for each objective (provided the test measures more than one objective and cut-off scores are necessary)?

N Y 46. Are instructions given for reporting and recording test results?

N Y 47. Has the possibility of special problems been taken into account?

N Y 48. Has the total test been demonstrated reliable by the calculation of ϕ for test-retest reliability (ϕ being greater than +.50)?

N Y 49. Did the sample used to check reliability consist of at least 50 people?

N Y 50. Was the sample used to check reliability selected randomly from the population of people who would ordinarily take this test?

N Y 51. Were "equivalent conditions" present for the test and the retest?

N Y 52. Were the trainees unaware that they would be tested again?

N Y 53. Were the tests given close together in time to eliminate learning or forgetting between testing?

N Y 54. Has the test been demonstrated valid through a content validity check?

N Y 55. Has the test been demonstrated valid through a concurrent validity check (ϕ being greater than +.50)?

N Y 56. Has the test been demonstrated valid through a predictive validity check (ϕ being greater than +.50)?

N Y 57. Are you thoroughly convinced that the test in question is suitable for administration?

APPENDIX C

GLOSSARY

Achievement Test - A test for measuring an individual's level of mastery of a subject. For example, an achievement test may be given on 4th grade mathematics to see if a student's mathematical ability has reached the 4th grade level. "Fourth grade level" may be defined in terms of the average 4th grader's scores, in which case the test would be norm-referenced, or in terms of math standards for 4th graders, in which case the achievement test would be criterion-referenced.

Aptitude Test - A test to determine an individual's learning capability in an area of instruction. For example, a test of mechanical aptitude would measure people's ability to learn to perform tasks involving mechanical skills and knowledges, not their present ability to perform mechanical tasks.

Conditions - One of the main parts of an objective that tells: 1) What the student has to work with, 2) the environmental circumstances under which the performance must be demonstrated, 3) what the student must work on, 4) his starting points, and 5) any limitations, special instructions, etc.

Course Criterion Test - A test given at the end of a course to determine if the student has reached the necessary criterion levels for the subject being taught. Course criterion tests are keyed to the course objectives and represent a "final exam" on meeting the standards specified in the objectives.

Criterion - Synonymous with standard (the part of the objective by which the performance is evaluated). For example, part of the criterion by which "donning a gas mask" is evaluated, is that the performance be completed in nine seconds or less. If it takes a trainee ten seconds to don the mask, he has not achieved the criterion level of performance.

Criterion-Referenced Test (CRT) - A CRT measures what an individual can do or knows, compared to what he must be able to do or must know in order to successfully perform a task. Here an individual's performance is compared to external criteria or performance standards which are derived from an analysis of what is required to do a particular task.

Critical Tasks - A task that if misperformed could lead to loss of life or property, or to mission failure. For example, in many first aid procedures, treating for shock is a critical task: Even if the other parts of the procedure are correctly performed, the individual may die of shock. Bandaging a wound, while important, would usually not be considered a critical task.

Diagnostic Test - A test used to inform a student of his progress, to determine if his behavior qualifies him for course entry, or to establish what objectives or steps he is weak on. For example, in BCT a diagnostic test is usually given before the comprehensive performance test (CPT)--thus, the student gets information on what he needs to improve before taking the CPT.

Entry Behavior - The performance of which a student is capable on a certain subject matter upon entering a course of instruction on that subject. Entry behavior may refer to skills, knowledges, and attitudes.

Error of Central Tendency - A rating error in which different raters tend to rate most students toward the middle of the scale. Thus, if there is a "neutral" point on a rating scale, raters may tend to rate most students close to it.

Error of Halo - A rating error made due to an observer being biased about an individual. This may be caused by an observer allowing his general impression of an individual to influence his judgment. The resulting shift of the rating can be toward the high end of the scale (positive halo) or the low end of the scale (negative halo).

Error of Standards - An error committed in rating due to differences in the observers' standards. One rater's standards might be higher than another rater's. Thus, while one rater might rate a person's performance as "unsatisfactory," another rater might rate that same person's performance as "satisfactory."

Fidelity - The extent to which a CRT resembles the actual objective (or performance) being tested. The more the CRT resembles the performance in question, the higher the fidelity of the CRT. For example, if you tested a person to see how well he could bandage a wound by observing him bandaging a wound, the test would have high fidelity. If you tested him by asking him to answer multiple-choice questions on how to bandage a wound, the test would have low fidelity.

Format - The type of test or item organization. Examples of item format include paper and pencil tests, hands-on performance tests, multiple choice tests, recall measures, job simulations, etc.

Hands-On Performance Measure - A type of performance measure where the individual is tested on the apparatus for which he was trained (no paper-and-pencil tests). A hands-on performance measure of generator repair would require the trainee to actually repair a generator.

Indicator - The action verb of the objective's task statement through which the ability to do the performance specified by the main intent is inferred, when the main intent itself is not directly observable. For example, if the main intent is "Discriminate between shears used for cutting a straight line in tin and those used for cutting a curved line," the indicator might be "by circling the picture of shears used for cutting a curved line." Note that in this case the main intent--"discriminate"--is covert; that is, it is not directly observable. Thus, an indicator had to be added.

Item Analysis - A technique used to help spot bad items. A number of techniques can be used to do this, all of which use the following principle: Acceptable items discriminate between "masters" and "non-masters." Unacceptable items are incapable of making such a discrimination. So, in item analysis, you look for items which are missed by "non-masters" and passed by "masters."

Item Pool - The total set of items constructed for a specified test, be it a single or multiple objective test. The item pool is reduced by item analysis and review techniques to yield a final version of the test consisting of the best items from the pool.

Learning Analysis - An analysis of the steps necessary to obtain the objective, the skills needed to learn the material presented, etc. In a learning analysis, you determine what skills, knowledge, and attitudes individuals must be taught to get them from their entry behaviors to the behaviors specified by the learning objectives.

Learning Objective - A learning objective describes what the individual must know and be able to do at the completion of training. It may be the same as a performance objective or may be less rigorous with respect to conditions and/or standards. Thus, a learning objective tells you what the individual should get out of training, not necessarily what he must be able to do on the job. An individual may require further training on the job after he has achieved a learning objective, before he is able to meet a performance objective. Learning objectives, like all objectives, have three main parts: performances (tasks), conditions, and standards.

Logical Error - An error in rating which may be due to an observer giving similar ratings to traits which aren't necessarily related. Two or more traits being rated at the same time may logically seem related to an observer when they really are not. For example, a rater might score a person similarly on "follows orders" and "completes work on time" because the two traits seem logically related, even though they are not necessarily related.

Main Intent - The statement of the task that tells you what the objective is mainly about: The skill or knowledge the learner is to develop, or the performance which is the purpose of the objective. A main intent may be overt (observable)--for example, "disassemble a M-16"; or covert (unobservable)--for example, "know the differences in appearance between poisonous and nonpoisonous snakes." If covert, an indicator must be added to the objective to tell you how to evaluate the main intent.

Mastery - An individual has attained mastery when he has completed the training segment that your CRT was developed to test and has passed the test, showing that he can perform at the minimal level necessary for successful task completion, or better.

Masters - People who are competent at performing a given task or who have already completed the training segment that a CRT is being developed to test. A master can perform the task(s) for which he has been trained.

Non-Masters - People who are not competent performers, or who are not knowledgeable in the subject matter being tested, or who have not had appropriate training.

Norm-Referenced Test (NRT) - An approach to testing in which an individual's test score is compared to the scores of other individuals regardless of standards specified by an objective.

Objective - A statement specifying skills and knowledge to be tested. It consists of three parts: 1) performance (task), 2) conditions, and 3) standards. Thus, an objective states what must be done (task), the conditions under which it must be done, and how well and/or how quickly it must be done (standards).

Percentile - A value on a scale of one hundred that indicates the percent of a distribution that is equal to or below it. For example, if a person scores at the 95th percentile, this means he has done better than 95 out of 100 people who have taken the test.

Performance - One of three main parts of an objective which states precisely what must be done. Every statement of performance includes an action verb. Sometimes this verb is the performance itself and sometimes it is an indicator of the performance.

Performance Measurement - The method used to ascertain whether or not an individual has achieved the specified criterion level on the performance of a particular task or tasks.

Performance Objective - A performance objective is derived from an analysis of what must be done in order to perform a task adequately. Like any objective, a performance objective has three main parts: performance (task), conditions, and standards. A performance objective is the highest level of objective--it tells what must be done in order to perform a task successfully.

Performance Tests - A performance test measures the individual's ability to perform a particular task or group of tasks. "Can he do the task properly or not?" is the question that a criterion-referenced performance test seeks to answer. A norm-referenced performance test investigates how well an individual can perform a task compared to other people. A performance test can be administered using actual hands-on performance, simulated performance, or in a paper-and-pencil format (if the performance in question requires use of paper-and-pencil--calculating azimuths, for example).

Phi Coefficient (ϕ) - A simple statistical technique which may be used for CRT item analysis if the following data are available: 1) which people pass which items, and 2) which people are "masters" and which are "non-masters."

$$\phi = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \quad \text{where}$$

A = number of "masters" who passed the item
B = number of "masters" who failed the item
C = number of "non-masters" who passed the item
D = number of "non-masters" who failed the item

ϕ may also be used as a measure of test-retest reliability and of concurrent or predictive validity. For such uses the formula remains the same, but the letters refer to different measures:

Test-Retest Reliability

		1st administration of test	
		Fail	Pass
2nd administration of test	Pass	B	A
	Fail	D	C

Concurrent or Predictive Validity

		CRT Results	
		Fail	Pass
Concurrent or predictive measure	Acceptable	B	A
	Unacceptable	D	C

Population - The universal set of individuals who possess the characteristic(s) in question. For example, the population possessing the characteristic "lives in the U.S.A." is the population of the U.S.A. The population of living U.S. citizens includes all people possessing U.S. citizenship whether or not they live in the U.S.A. The population possessing the characteristic "passed Army BCT during the last year" includes all Army personnel who have passed BCT in the last year.

Practical Constraints - Factors such as time availability, manpower availability, costs, etc. which may impair administration of test items if conditions and standards remain as presently specified in an objective. For example, an objective requiring the firing of nuclear projectiles may well have practical constraints--the objective would have to be modified so that the test item could substitute firing "dummy" nuclear projectiles.

Process Measurement - Measurement of a process rather than a product. Process measurement is indicated when an objective specifies a sequence of performances which can be observed and when the performances are as important as the final product of the performances. It is also appropriate when product cannot be distinguished from process or when the product cannot be measured for safety or other constraining reasons. Process measurement usually requires observing whether or not a performance is done properly and/or quickly enough, and in the right sequence. An example of process measurement is scoring a person "go" or "no-go" on his ability to properly execute an "about face" in drill and ceremonies.

Product Measurement - Measurement of a product rather than a process. Product measurement is appropriate if: 1) the objective specifies a product, 2) the product can be measured as to either presence or characteristics, and 3) the procedure leading to product can vary without affecting the product. An example of product measurement is observing a weapon to see if it has been reassembled correctly--here, you don't need to watch the weapon being reassembled (the process) because you can observe the product to see if it has been reassembled correctly.

Random Sample - A sample in which the individuals chosen from among all available people of the appropriate type are selected by chance. A random sample of a population would be composed of people possessing the characteristic of the population, each of whom is equally likely to be chosen from the population.

Rating Scale - A device used to evaluate achievement. When using a rating scale for scoring, you should specify the rating a student needs to achieve criterion level for the performance specified by the objective. A rating scale might also be used to assess entering behavior at the start of instruction. Rating scales usually have three to nine points on them representing levels of performance from low to high.

Reliability - Reliability is a synonym for "consistency" or "repeatability." A test is considered to be reliable if it makes the same discriminations among individuals on multiple occasions. People should score about the same each time they take the test, if it is reliable (assuming that they don't learn or forget between tests). Thus, a person's scores on reliable tests are consistent and repeatable.

Repertoire of Behavior - The group of behaviors which the student is capable of performing. Different groups have different repertoires of behaviors. For example, soldering connections is a part of the repertoire of behavior of electronic technicians, but probably not of food service specialists. Multiplying two single-digit numbers is part of the repertoire of behavior of many 10 year olds, but not of too many 7 year olds.

Representative Sample - A representative sample is one which reflects (represents) the population for which a test is intended. In order to try out test items on a representative sample, the persons in the sample should be similar to those for whom the test is intended. Thus, if a test is intended for people who have completed BCT, a representative sample would be composed of people who have completed BCT. If a test is intended for people who have completed a field wireman course, a representative sample whould be composed of people who have completed that course. If a population is sampled randomly, the resulting group will be a representative sample of that population--and not of any other population.

Screening Device - A device used to screen out trainees who do not qualify for the training course being considered, either because they are already masters of the subject matter or because they do not have the entry behavior required for the course. (A CRT can be used as a screening device.)

Simulation - A situation where phenomena likely to occur in actual performance can be reproduced under test conditions without using the real-life equipment. Simulation can use complex simulators--a simulated helicopter is an example--or simple simulators--a rubber bayonet is an example.

Skills - A learned ability to successfully perform a certain action or related group of actions. While knowledge is often necessary for skills, the knowledge of how to perform an act is not the skill--the performance of the act is the skill. Riding a bicycle, for example, is a skill requiring performance of a related sequence of actions. A person may have knowledge of how to ride--he could tell you how to sit, pedal, shift gears, brake, etc.--without possessing the skill of riding.

Standards - The third main part of an objective which specifies the criterion by which the performance is evaluated (how well and/or how quickly a performance must be done). There are several types of standards that may be included in any objective, any of which tell how well or how quickly the task must be done. An objective may have both a standard of quality and of speed.

Subject Matter Expert - Someone who is well qualified in the subject matter being tested. The reason for having such a person review the test items is because the test developer may not be an expert in the subject. A subject matter expert is usually trained and experienced in a particular subject area.

Task - A part of a job that requires certain performance(s). A group of tasks comprise a job, while complex tasks may be broken down into subtasks. The job of auto mechanic, for example, is composed of many tasks including tune-ups, repairing transmissions, replacing brake linings, etc. The task "tune-up" is composed of subtasks such as replace spark plugs, replace points, etc. The designation of tasks is often arbitrary. If, for example, a person's job was "tune-up specialist," replacing points would be a task rather than a subtask. Subtasks under "replacing points" would include removing old points, putting in new points, setting gap on new points, etc.

Task Analysis - An analysis of a task (or tasks) to determine the skills and knowledges necessary to perform it, equipment and/or facilities required, attitudes required, critical tasks, proper sequence of actions, etc. Sometimes, all the tasks in a given job are analyzed by a procedure called "job task analysis" or "job analysis." Often, task analysis is used as a synonym for job analysis.

Test Evaluation Unit - A group of people who are experts in the area of testing. Test evaluation personnel are often expert in educational technology--they can be of help with many training and testing problems.

Test-Retest Reliability - Determination of the stability of test scores by repeated testing. Test-retest reliability assumes that no training or forgetting takes place between test administrations, so both administrations should be given close together in time. If a test has high test-retest reliability, a person should score about the same each time he takes the test. If it has low test-retest reliability, a person's score may vary widely from one test administration to the next.

Validation - The process of determining whether a test actually measures what it is intended to measure.

Validity, Concurrent - Statements of concurrent validity indicate the extent to which a test may be used to estimate an individual's present standing on the criterion. This type of validity reflects only the status quo at a particular time. In concurrent validation, individuals' scores on the CRT are correlated with their performances on another measure of the objective(s) in question. If people who score high on the CRT score high on the other measure, while people who score low on the CRT score low on the other measure, the test is concurrently valid. Of course, the other measure must be a good one or the concurrent validation won't mean much.

Validity, Content - If test objectives are based on an adequate task analysis of what the individual must do, and if the test items measure exactly what the objectives say they should, the test is content valid. Content validation is especially appropriate for CRTs.

Validity, Predictive - Statements of predictive validity indicate the extent to which an individual's future level on a criterion can be predicted from a knowledge of his test performance. CRT scores are correlated with another measure of the same performance which is taken later, on the job. If high scores on the CRT are correlated with success on the job, while low scores are correlated with lack of success, the CRT has high predictive validity.

APPENDIX D

SQUARE ROOT TABLES

How To Use the Square Root Tables

For numbers 1 to 1,000: In column N, locate the number for which you want the square root, and immediately to the right, in Column $\sqrt{N}$, you will find the answer. For example, the square root of 150 is 12.2474.

For numbers 1,001 to 100,000: (1) Take the number for which you want the square root and move its decimal point two places to the left. (2) Round off to the nearest whole number, and find this number in Column N. (3) Take the number immediately to the right, in Column $\sqrt{N}$, and move its decimal point one place to the right. That is the square root.

For example, suppose you need the square root of 1,200. First, move the decimal point two places to the left. Since this gives you "12.00", no rounding is necessary. Then look up the square root of 12 in the square root table, and you find "3.46410". Then move the decimal point one place to the right and you have the answer: "34.6410."

In some cases, there will be slight rounding error, but this will not affect your computation of Ø. For example, using this procedure, you would find that the square root of 9,912 is 99.4987, when it is actually 99.5590. The difference--0.0603--is insignificant.

For numbers 100,001 to 10,000,000: (1) Take the number for which you want the square root and move its decimal point four places to the left. (2) Round off to the nearest whole number, and find this number in Column N. (3) Take the number immediately to the right, in Column $\sqrt{N}$, and move its decimal point two places to the right. That is the square root.

N	$\sqrt{N}$
1	1.00 000
2	1.41 421
3	1.73 205
4	2.00 000
5	2.23 607
6	2.44 949
7	2.64 575
8	2.82 843
9	3.00 000
10	3.16 228
11	3.31 662
12	3.46 410
13	3.60 555
14	3.74 166
15	3.87 298
16	4.00 000
17	4.12 311
18	4.24 264
19	4.35 890
20	4.47 214
21	4.58 258
22	4.69 042
23	4.79 583
24	4.89 898
25	5.00 000
26	5.09 902
27	5.19 615
28	5.29 150
29	5.38 516
30	5.47 723
31	5.56 776
32	5.65 685
33	5.74 456
34	5.83 095
35	5.91 608
36	6.00 000
37	6.08 276
38	6.16 441
39	6.24 500
40	6.32 456
41	6.40 312
42	6.48 074
43	6.55 744
44	6.63 325
45	6.70 820
46	6.78 233
47	6.85 565
48	6.92 820
49	7.00 000
50	7.07 107
N	$\sqrt{N}$

N	$\sqrt{N}$
50	7.07 107
51	7.14 143
52	7.21 110
53	7.28 011
54	7.34 847
55	7.41 620
56	7.48 331
57	7.54 983
58	7.61 577
59	7.68 115
60	7.74 597
61	7.81 025
62	7.87 401
63	7.93 725
64	8.00 000
65	8.06 226
66	8.12 404
67	8.18 535
68	8.24 621
69	8.30 662
70	8.36 660
71	8.42 615
72	8.48 528
73	8.54 400
74	8.60 233
75	8.66 025
76	8.71 780
77	8.77 496
78	8.83 176
79	8.88 819
80	8.94 427
81	9.00 000
82	9.05 539
83	9.11 043
84	9.16 515
85	9.21 954
86	9.27 362
87	9.32 738
88	9.38 083
89	9.43 398
90	9.48 683
91	9.53 939
92	9.59 166
93	9.64 365
94	9.69 536
95	9.74 679
96	9.79 796
97	9.84 886
98	9.89 949
99	9.94 987
100	10.00 000
N	$\sqrt{N}$

N	$\sqrt{N}$
100	10.00 00
101	10.04 99
102	10.09 95
103	10.14 89
104	10.19 80
105	10.24 70
106	10.29 56
107	10.34 41
108	10.39 23
109	10.44 03
110	10.48 81
111	10.53 57
112	10.58 30
113	10.63 01
114	10.67 71
115	10.72 38
116	10.77 03
117	10.81 67
118	10.86 28
119	10.90 87
120	10.95 45
121	11.00 00
122	11.04 54
123	11.09 05
124	11.13 55
125	11.18 03
126	11.22 50
127	11.26 94
128	11.31 37
129	11.35 78
130	11.40 18
131	11.44 55
132	11.48 91
133	11.53 26
134	11.57 58
135	11.61 90
136	11.66 19
137	11.70 47
138	11.74 73
139	11.78 98
140	11.83 22
141	11.87 43
142	11.91 64
143	11.95 83
144	12.00 00
145	12.04 16
146	12.08 30
147	12.12 44
148	12.16 55
149	12.20 66
150	12.24 74
N	$\sqrt{N}$

N	$\sqrt{N}$
150	12.24 74
151	12.28 82
152	12.32 88
153	12.36 93
154	12.40 97
155	12.44 99
156	12.49 00
157	12.53 00
158	12.56 98
159	12.60 95
160	12.64 91
161	12.68 86
162	12.72 79
163	12.76 71
164	12.80 62
165	12.84 52
166	12.88 41
167	12.92 28
168	12.96 15
169	13.00 00
170	13.03 84
171	13.07 67
172	13.11 49
173	13.15 29
174	13.19 09
175	13.22 88
176	13.26 65
177	13.30 41
178	13.34 17
179	13.37 91
180	13.41 64
181	13.45 36
182	13.49 07
183	13.52 77
184	13.56 47
185	13.60 15
186	13.63 82
187	13.67 48
188	13.71 13
189	13.74 77
190	13.78 40
191	13.82 03
192	13.85 64
193	13.89 24
194	13.92 84
195	13.96 42
196	14.00 00
197	14.03 57
198	14.07 12
199	14.10 67
200	14.14 21
N	$\sqrt{N}$

N	$\sqrt{N}$
200	14.14 21
201	14.17 74
202	14.21 27
203	14.24 78
204	14.28 29
205	14.31 78
206	14.35 27
207	14.38 75
208	14.42 22
209	14.45 68
210	14.49 14
211	14.52 58
212	14.56 02
213	14.59 45
214	14.62 87
215	14.66 29
216	14.69 69
217	14.73 09
218	14.76 48
219	14.79 86
220	14.83 24
221	14.86 61
222	14.89 97
223	14.93 32
224	14.96 66
225	15.00 00
226	15.03 33
227	15.06 65
228	15.09 97
229	15.13 27
230	15.16 58
231	15.19 87
232	15.23 15
233	15.26 43
234	15.29 71
235	15.32 97
236	15.36 23
237	15.39 48
238	15.42 72
239	15.45 96
240	15.49 19
241	15.52 42
242	15.55 63
243	15.58 85
244	15.62 05
245	15.65 25
246	15.68 44
247	15.71 62
248	15.74 80
249	15.77 97
250	15.81 14
N	$\sqrt{N}$

N	$\sqrt{N}$
250	15.81 14
251	15.84 30
252	15.87 45
253	15.90 60
254	15.93 74
255	15.96 87
256	16.00 00
257	16.03 12
258	16.06 24
259	16.09 35
260	16.12 45
261	16.15 55
262	16.18 64
263	16.21 73
264	16.24 81
265	16.27 88
266	16.30 95
267	16.34 01
268	16.37 07
269	16.40 12
270	16.43 17
271	16.46 21
272	16.49 24
273	16.52 27
274	16.55 29
275	16.58 31
276	16.61 32
277	16.64 33
278	16.67 33
279	16.70 33
280	16.73 32
281	16.76 31
282	16.79 29
283	16.82 26
284	16.85 23
285	16.88 19
286	16.91 15
287	16.94 11
288	16.97 06
289	17.00 00
290	17.02 94
291	17.0[illegible]87
292	17.08 80
293	17.11 72
294	17.14 64
295	17.17 56
296	17.20 47
297	17.23 37
298	17.26 27
299	17.29 16
300	17.32 05
N	$\sqrt{N}$

N	$\sqrt{N}$
300	17.32 05
301	17.34 94
302	17.37 81
303	17.40 69
304	17.43 56
305	17.46 42
306	17.49 29
307	17.52 14
308	17.54 99
309	17.57 84
310	17.60 68
311	17.63 52
312	17.66 36
313	17.69 18
314	17.72 00
315	17.74 82
316	17.77 64
317	17.80 45
318	17.83 26
319	17.86 06
320	17.88 85
321	17.91 65
322	17.94 44
323	17.97 22
324	18.00 00
325	18.02 78
326	18.05 55
327	18.08 31
328	18.11 08
329	18.13 84
330	18.16 59
331	18.19 34
332	18.22 09
333	18.24 83
334	18.27 57
335	18.30 30
336	18.33 03
337	18.35 76
338	18.38 48
339	18.41 20
340	18.43 91
341	18.46 62
342	18.49 32
343	18.52 03
344	18.54 72
345	18.57 42
346	18.60 11
347	18.62 79
348	18.65 48
349	18.68 15
350	18.70 83
N	$\sqrt{N}$

N	$\sqrt{N}$
350	18.70 83
351	18.73 50
352	18.76 17
353	18.78 83
354	18.81 49
355	18.84 14
356	18.86 80
357	18.89 44
358	18.92 09
359	18.94 73
360	18.97 37
361	19.00 00
362	19.02 63
363	19.05 26
364	19.07 88
365	19.10 50
366	19.13 11
367	19.15 72
368	19.18 33
369	19.20 94
370	19.23 54
371	19.26 14
372	19.28 73
373	19.31 32
374	19.33 91
375	19.36 49
376	19.39 07
377	19.41 65
378	19.44 22
379	19.46 79
380	19.49 36
381	19.51 92
382	19.54 48
383	19.57 04
384	19.59 59
385	19.62 14
386	19.64 69
387	19.67 23
388	19.69 77
389	19.72 31
390	19.74 84
391	19.77 37
392	19.79 90
393	19.82 42
394	19.84 94
395	19.87 46
396	19.89 97
397	19.92 49
398	19.94 99
399	19.97 50
400	20.00 00
N	$\sqrt{N}$

N	$\sqrt{N}$
400	20.00 00
401	20.02 50
402	20.04 99
403	20.07 49
404	20.09 98
405	20.12 46
406	20.14 94
407	20.17 42
408	20.19 90
409	20.22 37
410	20.24 85
411	20.27 31
412	20.29 78
413	20.32 24
414	20.34 70
415	20.37 15
416	20.39 61
417	20.42 06
418	20.44 50
419	20.46 95
420	20.49 39
421	20.51 83
422	20.54 26
423	20.56 70
424	20.59 13
425	20.61 55
426	20.63 98
427	20.66 40
428	20.68 82
429	20.71 23
430	20.73 64
431	20.76 05
432	20.78 46
433	20.80 87
434	20.83 27
435	20.85 67
436	20.88 06
437	20.90 45
438	20.92 84
439	20.95 23
440	20.97 62
441	21.00 00
442	21.02 38
443	21.04 76
444	21.07 13
445	21.09 50
446	21.11 87
447	21.14 24
448	21.16 60
449	21.18 96
450	21.21 32
N	$\sqrt{N}$

N	$\sqrt{N}$
450	21.21 32
451	21.23 68
452	21.26 03
453	21.28 38
454	21.30 73
455	21.33 07
456	21.35 42
457	21.37 76
458	21.40 09
459	21.42 43
460	21.44 76
461	21.47 09
462	21.49 42
463	21.51 74
464	21.54 07
465	21.56 39
466	21.58 70
467	21.61 02
468	21.63 33
469	21.65 64
470	21.67 95
471	21.70 25
472	21.72 56
473	21.74 86
474	21.77 15
475	21.79 45
476	21.81 74
477	21.84 03
478	21.86 32
479	21.88 61
480	21.90 89
481	21.93 17
482	21.95 45
483	21.97 73
484	22.00 00
485	22.02 27
486	22.04 54
487	22.06 81
488	22.09 07
489	22.11 33
490	22.13 59
491	22.15 85
492	22.18 11
493	22.20 36
494	22.22 61
495	22.24 86
496	22.27 11
497	22.29 35
498	22.31 59
499	22.33 83
500	22.36 07
N	$\sqrt{N}$

N	$\sqrt{N}$
500	22.36 07
501	22.38 30
502	22.40 54
503	22.42 77
504	22.44 99
505	22.47 22
506	22.49 44
507	22.51 67
508	22.53 89
509	22.56 10
510	22.58 32
511	22.60 53
512	22.62 74
513	22.64 95
514	22.67 16
515	22.69 36
516	22.71 56
517	22.73 76
518	22.75 96
519	22.78 16
520	22.80 35
521	22.82 54
522	22.84 73
523	22.86 92
524	22.89 10
525	22.91 29
526	22.93 47
527	22.95 65
528	22.97 83
529	23.00 00
530	23.02 17
531	23.04 34
532	23.06 51
533	23.08 68
534	23.10 84
535	23.13 01
536	23.15 17
537	23.17 33
538	23.19 48
539	23.21 64
540	23.23 79
541	23.25 94
542	23.28 09
543	23.30 24
544	23.32 38
545	23.34 52
546	23.36 66
547	23.38 80
548	23.40 94
549	23.43 07
550	23.45 21
N	$\sqrt{N}$

N	$\sqrt{N}$
550	23.45 21
551	23.47 34
552	23.49 47
553	23.51 60
554	23.53 72
555	23.55 84
556	23.57 97
557	23.60 08
558	23.62 20
559	23.64 32
560	23.66 43
561	23.68 54
562	23.70 65
563	23.72 76
564	23.74 87
565	23.76 97
566	23.79 08
567	23.81 18
568	23.83 28
569	23.85 37
570	23.87 47
571	23.89 56
572	23.91 65
573	23.93 74
574	23.95 83
575	23.97 92
576	24.00 00
577	24.02 08
578	24.04 16
579	24.06 24
580	24.08 32
581	24.10 39
582	24.12 47
583	24.14 54
584	24.16 61
585	24.18 68
586	24.20 74
587	24.22 81
588	24.24 87
589	24.26 93
590	24.28 99
591	24.31 05
592	24.33 11
593	24.35 16
594	24.37 21
595	24.39 26
596	24.41 31
597	24.43 36
598	24.45 40
599	24.47 45
600	24.49 49
N	$\sqrt{N}$

N	$\sqrt{N}$
600	24.49 49
601	24.51 53
602	24.53 57
603	24.55 61
604	24.57 64
605	24.59 67
606	24.61 71
607	24.63 74
608	24.65 77
609	24.67 79
610	24.69 82
611	24.71 84
612	24.73 86
613	24.75 88
614	24.77 90
615	24.79 92
616	24.81 93
617	24.83 95
618	24.85 96
619	24.87 97
620	24.89 98
621	24.91 99
622	24.93 99
623	24.96 00
624	24.98 00
625	25.00 00
626	25.02 00
627	25.04 00
628	25.05 99
629	25.07 99
630	25.09 98
631	25.11 97
632	25.13 96
633	25.15 95
634	25.17 94
635	25.19 92
636	25.21 90
637	25.23 89
638	25.25 87
639	25.27 84
640	25.29 82
641	25.31 80
642	25.33 77
643	25.35 74
644	25.37 72
645	25.39 69
646	25.41 65
647	25.43 62
648	25.45 58
649	25.47 55
650	25.49 51
N	$\sqrt{N}$

N	$\sqrt{N}$
650	25.49 51
651	25.51 47
652	25.53 43
653	25.55 39
654	25.57 34
655	25.59 30
656	25.61 25
657	25.63 20
658	25.65 15
659	25.67 10
660	25.69 05
661	25.70 99
662	25.72 94
663	25.74 88
664	25.76 82
665	25.78 76
666	25.80 70
667	25.82 63
668	25.84 57
669	25.86 50
670	25.88 44
671	25.90 37
672	25.92 30
673	25.94 22
674	25.96 15
675	25.98 08
676	26.00 00
677	26.01 92
678	26.03 84
679	26.05 76
680	26.07 68
681	26.09 60
682	26.11 51
683	26.13 43
684	26.15 34
685	26.17 25
686	26.19 16
687	26.21 07
688	26.22 98
689	26.24 88
690	26.26 79
691	26.28 69
692	26.30 59
693	26.32 49
694	26.34 39
695	26.36 29
696	26.38 18
697	26.40 08
698	26.41 97
699	26.43 86
700	26.45 75
N	$\sqrt{N}$

N	$\sqrt{N}$
700	26.45 75
701	26.47 64
702	26.49 53
703	26.51 41
704	26.53 30
705	26.55 18
706	26.57 07
707	26.58 95
708	26.60 83
709	26.62 71
710	26.64 58
711	26.66 46
712	26.68 33
713	26.70 21
714	26.72 08
715	26.73 95
716	26.75 82
717	26.77 69
718	26.79 55
719	26.81 42
720	26.83 28
721	26.85 14
722	26.87 01
723	26.88 87
724	26.90 72
725	26.92 58
726	26.94 44
727	26.96 29
728	26.98 15
729	27.00 00
730	27.01 85
731	27.03 70
732	27.05 55
733	27.07 40
734	27.09 24
735	27.11 09
736	27.12 93
737	27.14 77
738	27.16 62
739	27.18 46
740	27.20 29
741	27.22 13
742	27.23 97
743	27.25 80
744	27.27 64
745	27.29 47
746	27.31 30
747	27.33 13
748	27.34 96
749	27.36 79
750	27.38 61
N	$\sqrt{N}$

N	$\sqrt{N}$
750	27.38 61
751	27.40 44
752	27.42 26
753	27.44 08
754	27.45 91
755	27.47 73
756	27.49 55
757	27.51 36
758	27.53 18
759	27.55 00
760	27.56 81
761	27.58 62
762	27.60 43
763	27.62 25
764	27.64 05
765	27.65 86
766	27.67 67
767	27.69 48
768	27.71 28
769	27.73 08
770	27.74 89
771	27.76 69
772	27.78 49
773	27.80 29
774	27.82 09
775	27.83 88
776	27.85 68
777	27.87 47
778	27.89 27
779	27.91 06
780	27.92 85
781	27.94 64
782	27.96 43
783	27.98 21
784	28.00 00
785	28.01 79
786	28.03 57
787	28.05 35
788	28.07 13
789	28.08 91
790	28.10 69
791	28.12 47
792	28.14 25
793	28.16 03
794	28.17 80
795	28.19 57
796	28.21 35
797	28.23 12
798	28.24 89
799	28.26 66
800	28.28 43
N	$\sqrt{N}$

N	$\sqrt{N}$
800	28.28 43
801	28.30 19
802	28.31 96
803	28.33 73
804	28.35 49
805	28.37 25
806	28.39 01
807	28.40 77
808	28.42 53
809	28.44 29
810	28.46 05
811	28.47 81
812	28.49 56
813	28.51 32
814	28.53 07
815	28.54 82
816	28.56 57
817	28.58 32
818	28.60 07
819	28.61 82
820	28.63 56
821	28.65 31
822	28.67 05
823	28.68 80
824	28.70 54
825	28.72 28
826	28.74 02
827	28.75 76
828	28.77 50
829	28.79 24
830	28.80 97
831	28.82 71
832	28.84 44
833	28.86 17
834	28.87 91
835	28.89 64
836	28.91 37
837	28.93 10
838	28.94 82
839	28.96 55
840	28.98 28
841	29.00 00
842	29.01 72
843	29.03 45
844	29.05 17
845	29.06 89
846	29.08 61
847	29.10 33
848	29.12 04
849	29.13 76
850	29.15 48
N	$\sqrt{N}$

N	$\sqrt{N}$
850	29.15 48
851	29.17 19
852	29.18 90
853	29.20 62
854	29.22 33
855	29.24 04
856	29.25 75
857	29.27 46
858	29.29 16
859	29.30 87
860	29.32 58
861	29.34 28
862	29.35 98
863	29.37 69
864	29.39 39
865	29.41 09
866	29.42 79
867	29.44 49
868	29.46 18
869	29.47 88
870	29.49 58
871	29.51 27
872	29.52 96
873	29.54 66
874	29.56 35
875	29.58 04
876	29.59 73
877	29.61 42
878	29.63 11
879	29.64 79
880	29.66 48
881	29.68 16
882	29.69 85
883	29.71 53
884	29.73 21
885	29.74 89
886	29.76 58
887	29.78 25
888	29.79 93
889	29.81 61
890	29.83 29
891	29.84 96
892	29.86 64
893	29.88 31
894	29.89 98
895	29.91 66
896	29.93 33
897	29.95 00
898	29.96 66
899	29.98 33
900	30.00 00
N	$\sqrt{N}$

N	$\sqrt{N}$
900	30.00 00
901	30.01 67
902	30.03 33
903	30.05 00
904	30.06 66
905	30.08 32
906	30.09 98
907	30.11 64
908	30.13 30
909	30.14 96
910	30.16 62
911	30.18 28
912	30.19 93
913	30.21 59
914	30.23 24
915	30.24 90
916	30.26 55
917	30.28 20
918	30.29 85
919	30.31 50
920	30.33 15
921	30.34 80
922	30.36 45
923	30.38 09
924	30.39 74
925	30.41 38
926	30.43 02
927	30.44 67
928	30.46 31
929	30.47 95
930	30.49 59
931	30.51 23
932	30.52 87
933	30.54 50
934	30.56 14
935	30.57 78
936	30.59 41
937	30.61 05
938	30.62 68
939	30.64 31
940	30.65 94
941	30.67 57
942	30.69 20
943	30.70 83
944	30.72 46
945	30.74 09
946	30.75 71
947	30.77 34
948	30.78 96
949	30.80 58
950	30.82 21
N	$\sqrt{N}$

N	$\sqrt{N}$
950	30.82 21
951	30.83 83
952	30.85 45
953	30.87 07
954	30.88 69
955	30.90 31
956	30.91 92
957	30.93 54
958	30.95 16
959	30.96 77
960	30.98 39
961	31.00 00
962	31.01 61
963	31.03 22
964	31.04 83
965	31.06 44
966	31.08 05
967	31.09 66
968	31.11 27
969	31.12 88
970	31.14 48
971	31.16 09
972	31.17 69
973	31.19 29
974	31.20 90
975	31.22 50
976	31.24 10
977	31.25 70
978	31.27 30
979	31.28 90
980	31.30 50
981	31.32 09
982	31.33 69
983	31.35 28
984	31.36 88
985	31.38 47
986	31.40 06
987	31.41 66
988	31.43 25
989	31.44 84
990	31.46 43
991	31.48 02
992	31.49 60
993	31.51 19
994	31.52 78
995	31.54 36
996	31.55 95
997	31.57 53
998	31.59 11
999	31.60 70
1000	31.62 28
N	$\sqrt{N}$

APPENDIX E

REVIEW QUESTIONS AND ANSWERS

Frederick Steinheiser, Jr.
U.S. Army Research Institute for the Behavioral and Social Sciences

This Appendix contains a set of questions and answers for each chapter. This is not a set of test items. Rather, it is suggested that you attempt to answer each question for a given chapter after reading that chapter. You can then check your answer with the supplied answer.

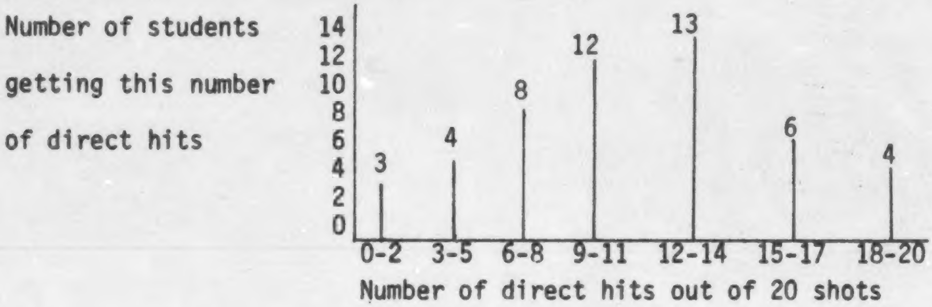
In many instances, the questions and answers supplement the material provided in the chapter. Hence, it will be a "learning experience" for you to study these questions and answers. A few questions were designed to be thought-provoking, and will require some creative insight and application of the information furnished in the text.

REVIEW PROBLEMS FOR CRT MANUAL

Chapter 1

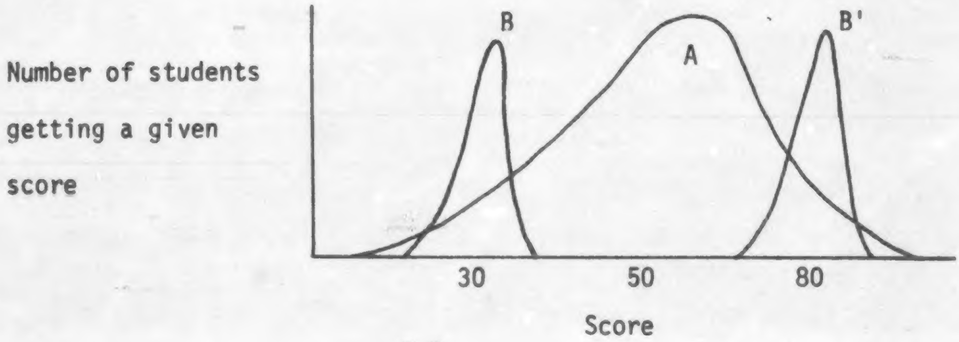
1. One of the important differences between norm-referenced tests and criterion-referenced tests is this: an NRT has mostly knowledge-type items, whereas a CRT has mainly performance-type items. (For example, writing down the steps in cleaning an M-16 vs. actually cleaning it properly.) True or false?

2. 50 students went to the rifle range, and each shot 20 rounds. The spread of scores looked like this:



To help you in reading this graph, note that 4 students scored from 3 to 5 direct hits. The instructor decided after the exercise to exempt the top 20% of the students from further practice, while the bottom 80% had to stay for more drill. How many students had to stay for more practice? Is this marksmanship test an example of a CRT or NRT, based upon the instructor's scoring procedure?

3. It's often helpful to plot a graph of test data, in order to get a visual impression of the distribution of scores. The distribution from an NRT is often quite different from the one of a CRT. (a) In the distributions below, which one(s) do you think came from an NRT, and which from a CRT? (b) The three scores of 30, 50, 80, shown below, tell different stories, depending upon whether they relate to the NRT or CRT distribution(s). How might you interpret these scores? (c) What are some possible reasons (think about both training and testing) for the differences in the shapes of the CR and NR scores as shown?



4. In comparing a large number of scores on a CRT before and after training, the CRT is being used (a) as a diagnostic aid, (b) to evaluate the instructor or program of instruction, (c) as a screening device.

5. A student got 90% of the problems on a math test correct, so he was advanced directly to the computer course without having to take a math refresher course. This math CRT was used (a) as a diagnostic aid, (b) to evaluate the instructor or program of instruction (c) as a screening device.

6. A student passed every item on a test except one. He was then allowed to enter the instruction program at the level of the test item that he missed. The information from this CRT was used (a) as a diagnostic aid, (b) to evaluate the instructor or program of instruction, (c) as a screening device.

Chapter 2

1. Hitting the outline of a moving enemy tank with an anti-tank round is an example of a level one, level two, or level three objective?

2. Hitting an enemy tank in actual combat with an anti-tank round is an example of a level one, level two, or level three objective?

3. Hitting the bull's eye of a stationary circular target with an anti-tank round is an example of a level one, two, or three objective?

4. It is possible that a poorly specified test item given after one phase of training might really be properly specified if given after another phase of training. True or false? (Hint: Think of the type of instructions or information given to solve a problem in an introductory vs. an intermediate course.)

5. Matching. Match each example with the appropriate technical term. The most significant parts of some examples are underlined.

   a. Performance   b. Conditions   c. Standards

   1. An action verb tells what is to be done by the student.
   2. The task must be performed to a satisfactory criterion level.
   3. The dial setting must be correct, to the nearest 1/2 degree.
   4. A student has to tune a jeep engine using only the tools provided.
   5. An indicator is essential in order to measure the main intent.
   6. Just because a student can pass a hands-on test in the classroom does not guarantee that he'll be able to pass the same test in simulated (or real) combat.

6. The use of "unitary objectives" (a) requires that all tasks be independent, (b) is the implementation of a Level One objective (but not Level Two or Three), (c) means that you don't have to divide objectives into Performance, Conditions, and Standards, (d) requires performance on more than one task at a time. (More than one choice may be correct.)

7. "Given these pictures of five tools, identify the one used for removing spark plugs by circling it." What is the main intent of the objective? What is the indicator? What are some other indicators that could be used without changing the main intent or the conditions?

8. "Cut a 6 inch diameter circle out of this piece of sheet metal using the appropriate shears." What is the main intent of this objective? What is the indicator?

9. Why is it essential that covert main intents have appropriate indicators?

10. In a couple of sentences, explain what is meant by specifying performances, conditions, and standards in "clear, operational terms."

11. Conditions and standards as specified for a Level One objective may actually be improperly specified for a Level Two objective. True or false?

12. Here's an extra "thought problem:"
Suppose that an instructor decided to test a helicopter pilot trainee without reference to explicit objectives. He merely "went along for the ride" while the student executed various maneuvers of his own choosing, and without knowing exactly which ones he ought to do or what the passing criterion was. (This is, of course, a highly unrealistic example, but it will help to focus upon some very realistic issues that crop up in the use of criterion referenced tests.)
After studying this CRT manual, the instructor thought that he would be able to improve his test. How might he go about it? (You don't have to be an expert in helicopter terminology to come up with a few overall suggestions.) What kinds of data might the instructor want to record when the student is executing various maneuvers?

Chapter 3

1. Giving a trainee a paper and pencil test on how to fire a mortar is of higher fidelity than evaluating him on a dry-fire test. True or false?

2. At the end of a medic's training, the instructor decided to pass only those students who got at least 40 out of 50 paper and pencil test items correct. Do you think that this was a good type of test to certify a student as a medic? Why or why not? How would you improve the test?

3. Another medical instructor decided to give his students 30 simulated injuries on dummies to treat, out of the total of 40 such injuries that had been covered in the course. A passing score was 25 out of the 30 injuries had to be treated perfectly. How does this test compare to the first instructor's test? What might be done to improve upon this test?

4. Another medical instructor gave his students all 40 of the injuries that had been taught in the course on the test dummies. A passing score was 38 out of 40. How does this test compare to the first two tests mentioned above? What might still be done to improve this test, assuming that no practical constraints stood in the way? What if there were constraints, so that not all students could be tested on all the injuries?

5. Which is not an "objective" test: (a) true-false, (b) matching, (c) essay, (d) multiple choice, (e) completion or fill-in-the-blank.

6. Having a person conduct the testing who was not the course instructor may help to eliminate the error of (a) standards, (b) logic, (c) central tendency, (d) halo.

7. Match the type of measurement with the correct example.

   a. Process   b. Product   c. Process and Product

   1. Find out if this battery has enough charge to start a jeep.
   2. Using dry-fire techniques, fire 10 M-102 Howitzer rounds for these ten target settings.
   3. Using the proper procedures during live fire for the above howitzer, at least 5 out of 10 rounds must impact within 25 meters of the target.

8. What are some general reasons that may make it necessary to modify conditions and standards from an ideal to a more practical setting?

9. Item sampling within objectives (a) is used where a concept must be learned, (b) is used where there is a routine process to be learned, (c) requires that a number of similar test items be produced from the total (possibly infinite) number of such items, (d) means that the same objective should be tested using a number of different items, (e) means that the same items are derived from different objectives. (More than one choice may be correct.)

10. Why should both easy and difficult conditions be used when testing under multiple conditions?

11. Sgt. Smith suspects that PFC Jones may not really be able to remove the spark plugs in one minute or less. Jones' times for three spark plugs were 59, 58, 58 sec. The next lowest score was by Duncan, whose times were 50, 52, and 53 sec. So Sgt. Smith singled out PFC Jones to do a fourth plug removal, as an extra (and unplanned) part of the test. Do you agree with Smith's decision? Why or why not?

12. How many decision points are there in the flow chart on p. 35?

Chapter 4

1. What are the specific steps of the Test Plan Worksheet? How are they to be used?

2. Evaluate this statement: "Good instructions do not give any hints to the students. The more that a student taking a test has to figure out for himself about the test, the better the test."

3. An inadequate test item is one which (a) is of low fidelity, (b) requires an indicator response, (c) is of high fidelity, (d) has stricter conditions than those which were stated in the objective, (e) has good agreement between the standards of the objective and the test item.

hapter 5

1. In choosing a group of Non-Masters, why can't you just choose people from any group which has not had the training experience that your group of Masters has had?

2. An instructor was designing a new electronics course. He decided that he needed 40 items on his final exam. On how many people should he try out this version of the exam? How many should be Masters, and how many should be Non-Masters?

3. Continuing with the above example, question #4 on this try-out exam was multiple choice, dealing with the voltage drop in a step-down transformer; 26 of the recent grads chose the correct answer, whereas 6 of the non-masters selected it. What do you think about the value of this item?

4. Question #17 was a true-false item, asking if a tunnel diode could be substituted for a malfunctioning capacitor if wired in parallel to the nearest transistor; 18 of the recent grads got it right, whereas 13 of the non-masters got it right. What do you think about the value of this item?

5. Question #14 asked if household voltage was a.c. or d.c.; 30 of the grads got it right, and 29 of the non-masters got it right. What do you think about the value of this item?

Chapter 6

For each of the terms discussed in this chapter, select the appropraite example or description. There are no duplications.

a. Personal Variables
b. Scoring
c. Fixed Point
d. Go/No-Go
e. Hands-On
f. False Positive
g. Rating Scale
h. Familiarization
i. False Negative
j. Assist Scoring
k. Uniform Instructions
l Environmental Variable

1. On Monday, PFC Jones passed a practice test, which his instructor said was just like the real one that was to be given on Wed. But Jones caught the flu on Tuesday, and still took the test on Wed. He failed the test, and as a result was not graduated into the next sequence of instruction.
2. All students should be equally alert, not hungry or tired.
3. Tester should know how to give the test, perhaps by having watched someone else conduct it previously.
4. Testing with the real device, apparatus, weapon, or machine.
5. The student has to do only those items again which he missed, and does not have to retake the whole test.
6. Student either knows how, or doesn't know how, there's no in-between "partial knowledge."
7. Conditions that, if changed from one group to the next, might (falsely) suggest that there's something wrong or unreliable about the test.
8. Numbers are assigned to performance on each item.
9. If a numerical answer is close enough to the correct anwer, it will be scored as correct.
10. Don't give extra hints or play favorites with people taking the test.
11. Determine if the student's performance met the specified standard.
12. PFC Smith has just advanced from the introductory to the intermediate automotive repair course. He was not able to tune and engine completely at the start of the intermediate course--although he had done so in order to pass the introductory course.
13. Although a student mechanic successfully passed the engine tuning section of an automotive CRT, he lost 1 tool, broke another, and got grease all over the place. Is this aspect of his performance significant, although it was not explicitly "tested" by any items of the actual test?

14. If a student passes (a) 2, (b) 3, (c) 4 objectives on a CRT with 4 objectives, then he should be passed on the whole test.

Chapter 7

1. "Reliability," when talking about tests, means about the same as (a) validity, (b) that the same scores should obtain on a second administration of the test to the same people, (c) that the test measures what it's supposed to measure, (d) standardization of training and testing conditions.

2. If validity is high, reliability will usually be (a) high, (b) low, (c) could be either high or low.

3. A test could be very reliable but not very valid. True or false? Can you think of an example to back up your answer?

4. Higher fidelity test items may help to increase (a) reliability, (b) validity, (c) both, (d) neither.

5. Why should only a short time (like a couple of days) elapse when conducting a test and retest reliability check?

6. A class of 30 M.P. students took a test at 1000 on Monday, and were given the same test (because the instructor wanted to conduct a reliability check) on Tuesday at 1900. (1900 was the only time that he could get all of the students together.) The results were:

		First Day	
		Fail	Pass
Second Day	Pass	2	17
	Fail	1	10

Compute the value of phi. What does this value suggest?

7. Another instructor decided to compare the results of his CRT given to the 28 students in his class with ratings of each student's performance as given by an expert observer. The results were:

		CRT Results	
		Fail	Pass
Expert's Ratings	Pass	1	20
	Fail	5	2

Compute the value of phi. What does this value suggest?

8. Here's another "thought question" that will help to prepare you for some of the more complex uses of CRTs in operational situations. A Corps of Engineers test produced the following results:

		Form A given on Mon.	
		Fail	Pass
Form A given on Wed.	Pass	5	22
	Fail	2	11

What is the value of phi, for test-retest reliability? Is it an acceptable value?
The Instructor was not pleased with this value of phi, and so he gave the same class another form of the test (Form B) on Fri. His aim was to compare the results from Form B with the results of Form A, as the latter was given on Mon. and Wed. The new data looked as follows:

		Form A on Mon.	
		Fail	Pass
Form B on Friday	Pass	1	35
	Fail	3	1

		Form A on Wed.	
		Fail	Pass
Form B on Friday	Pass	6	28
	Fail	2	5

What are the values of phi for these two tables?
Now interpret the values of all three coefficients that you've calculated; that is, what do you think the phi values for Form A on Mon. vs. Form B, and Form A on Wed. vs. Form B mean?

ANSWERS TO REVIEW PROBLEMS

Chapter 1

1. False. Review page 1-2. And the important differences between NRTs and CRTs are listed in Fig. 1-1.

2. If the standard specified in this problem is used, then 40 students will have to stay for more practice. This is an NRT, because the tester chose a passing standard on the basis of how well a student performed relative to other students. Note that with this kind of decision standard, only the top 20% of the students would pass even if (a) all students had performed "poorly" (all had obtained only 7 or less direct hits), or (b) all students had performed "very well" (all obtained 15 or more direct hits).

3a. Distribution A is from an NRT, whereas B and B' are from a CRT.

3b. Score of 80--on the NRT, only a small percentage of the students got this score or higher; on the CRT, most of the people whom we might label "master" got a score near 80.
Score of 50--on the NRT, more people got this score than any other score; whereas on the CRT, no one got this middle score.
Score of 30--on the NRT, only a small percentage of the students got this score or lower; whereas on the CRT, most of the people whom we might label as "non-masters" got a score near 30.
The NRT spreads people out on a distribution of scores, so that very few students do really well on the test, and very few do really poorly. Most tend to cluster around the middle, or average. The CRT ideally tries to spread people into two separate and non-overlapping groups: those who clearly passed the test, and those who clearly failed to pass it. (Masters and non-masters, or distributions B and B'.)

3c. There may be several reasons for the differences in the shapes of the curves. Consider differences in training procedures. Students described by curve A (the NR curve) may have been trained in a group, and given the same amount of training before being tested. Students described by curve B' may have received individually prescribed instruction (each student learning at his own pace), and then tested when he felt prepared to take the test.
Note that an NRT is designed to spread people out at the extreme scores, so that very few people do really well, and very few people do really poorly. Most people fall near the middle. A CRT is designed so that people who really have mastered the material will do well, and those who have not will do poorly on the test. A CRT is not used to assign grades to people, other than "pass-fail." If we use a CRT, we must care more about whether person X has mastered the task than if person X got a better score than person Y.

Consider, as a simple example, the "task" of broad-jumping. If we measure how far each person can jump, then we're using the distance measurement as an NRT. As a result of these measurements, we'll know if person X can jump farther than person Y, and we'll be able to plot a distribution of scores as in distribution A. Now suppose that we dig a 1.5 meter ditch, as the minimum criterion distance that a person must be able to jump in order to pass the jumping test. If a person can jump the ditch, we'll pass him; if not he'll fall in, and it will be obvious that he failed. This CRT is pass-fail oriented, since we're not interested in how far each student jumped. Rather, we just want to know if each student was able to jump across the ditch.

4. b.

5. c.

6. a.

Chapter 2

1. Level Two. This is a very close approximation ("high fidelity") to the "real world" situation.

2. Level One. This is the "real world" situation, which is impossible to totally duplicate in any kind of test setting.

3. Level Three. The target used here is much more artificial than the outline of moving tank, which we just described as a Level Two objective. In general, Level Three objectives must be passed before Level Two objectives are tested. Obviously, a student must learn how to load and fire an anti-tank round before he can even hope to hit the center of a stationary target.
   What level objective would this learning process be? Also a level Three. Piecemeal assessment of a subcomponent of the actual desired behavior in an artificial setting constitutes a Level Three Objective. So this example actually involved only two Level Three objectives: making sure that the weapon can be loaded and fired correctly, and then testing the student's accuracy of firing at an "artificial" target.

4. True. For example, a student at the end of a training sequence should not need the broad hints that you gave him during the earlier phases of training. Thus, early in an electronics course the test conditions might specify the specific components or instruments to be used in trouble-shooting malfunctioning equipment.

5. 1-a. 2-c. 3-c. 4-b. 5-a. 6-b.

6. a, d.

7. Main intent: Identify or recognize the spark plug wrench. Indicator: circling the picture of the wrench. Alternative indicators: Pointing out the picture, or placing a check mark by the picture.

8. The student has to first choose the appropriate shears, and then use them properly in order to cut a six inch circle. So the first main intent is the actual choice of the correct tool; the second (and perhaps more important) main intent is the correct use of the tool in cutting the sheet metal.

9. Overt (think of "open") main intents specify the required performance, tell how to measure it, and do not require indicator responses. Covert (think of "covered") main intents do not allow us to directly measure the desired performance. For example, an anti-aircraft test might require the gunnery crew to distinguish between the outlines of friendly vs. hostile planes. One way to conduct the test would be to have gunnery students draw pictures of Phantoms, MIGs, etc. A simpler and better indicator would be to give black profiles of all such aircraft, and have the student indicate (by circling, placing a checkmark, etc.) whether each craft is friendly or hostile.

10. Performances should be stated by specific action verbs. Conditions and standards will not be adequate if you have to supply any additional information. You should not have to interpret or figure out what is meant by the conditions and standards of statements if they are operationally defined.

11. True. Recall that a Level One objective refers to actual objectives in meaningful units of work activity in operational environments; "on-the-job-performance."
On the other hand, Level Three objectives include enabling skills and learning elements. A person must be able to perform these in order to correctly perform Level Two and One objectives. As an example, a Level One conditions statement might be: "Given a malfunctioning generator..." This would be appropriate for testing an advanced electrical technician, but not for one who had just completed the beginning course. The more appropriate conditions statement for the novice student should include more specific information ("helpful hints"), such as: "Given as 45 KW generator with a broken shaft bearing..." This would then be a Level Two (or even Three) conditions statement.
This example shows that improperly specified conditions at one level of objective may indeed be properly specified at another level.

12. Consider how the instructor could increase the structure and specificity of testing. How? By setting various objectives: Performance (handling the proper controls in the right sequence), Conditions (executing different maneuvers, flying with or against the wind, with and without a couple of tons of dead weight), and Standards (landing on a given target, making a "soft" landing, etc.). He should have a checklist of these many objectives made up before testing the trainee, so that he won't have to rely on his own intuitive evaluation and memory for what the entire set of scores was.
The instructor would want to record such data as: errors that the student made in carrying out various maneuvers, student's response times and hesitation, whether the student's response brought the craft within the range of the appropriate standard (did he fly on course, did he land on target, etc?).

Chapter 3

1. False. Higher fidelity items are more realistic and require "hands-on" performance.

2. No. This is only a paper and pencil test. You should have the trainees perform some of the behaviors that they will be required to perform on the job. Getting only 40 out of 50 questions correct also seems to be a rather lax standard, especially in a critical area like medical training. Incomplete or imperfect knowledge could result in needless suffering or even death.

3. This is better, because it is now a simulated "hands-on" performance test. However, only 30 test items (out of the 40 injuries which had been covered in the course) have been chosen from the 40 cases studied in the course. And only 25 of the 30 items need to be passed. So this less-than-full coverage also seems to be a rather lax standard.

4. This is a better test. Assuming that the items were reliable and valid (see chapters 5 and 7), the only obvious way to improve the test would be to increase the number of items. This would cover more variations of the original 40 types of injuries. If there were practical constraints as proposed, you might then want to randomly divide the class into two groups of 25 students each. Then randomly divide the 40 test items into two groups of 20 each. Thus, each student would get only 20 problems, but he would not know which 20 beforehand. He would have to do all 20 correctly.

5. c,e. All of the other choices in this answer could be "machine-scored." Be aware that sometimes more than one answer can be correct in fill-in-the blank items. Both this type of an item, and essay questions require judgment by the scorer.

6. d. The instructor might be tempted to give his own students slightly higher marks just to make himself look good.

7. 1-b. 2-a. (Only the settings are measured--no livefire is used.) 3-c.

8. You may have to cut down on the amount of supplies used in the test: fuel, ammunition, etc., because of excessive cost. You may have to conduct the test for a shorter time length than you'd like to, because of: large numbers of students, small number of judges, limited availability of test site.

9. a, c, d.

10. Suppose that the subject fails under one or more of the difficult conditions. Was it because he couldn't do the *task* at all, or because a *condition* was just too difficult? If you have one easy condition, and the subject passes that phase of the test, you'll at least know that he can do the task, although perhaps *not* under *all conditions* of difficulty.

11. No. He's letting his own subjective feelings and perhaps personal dislike bias his interpretation of the scores for Jones. "It is never proper to add test items during a test administration (p. 3-31)."

12. Five. Each of the "diamonds" requires that a yes-no decision be made at that point.

Chapter 4

1. The column headings in Fig. 3-11 indicate the specific guidelines which are explained in more detail on p. 4-2. In actual practice, it may often be easiest if you first of all make up a test item from your own assessment of the guidelines, and then check it against the specifications listed in Fig. 3-11. That is, after you've created a test item and specified the performance, conditions, and standards, all you have left to do is fill in the columns of the worksheet.

2. Note that on p. 4-6, hints are acceptable. Furthermore, the guidelines on p. 4-7 suggest that as a general rule, specific instructions should be supplied to the student. Hands-on performance items should have performance, conditions, and standards explicitly stated in operational terms.

3. d. Performance, conditions, and standards must match in the objective and in the test item. Level of fidelity, by itself, does not make an item good or bad. And an objective may have an overt main intent or require an indicator response.

1. The non-masters group must be composed of people who have met the minimal requirements for entering the course. They should be an actual sample of, or at least represent the people who will be taking the course. Think of how absurd it would be to use as the non-masters a group of secretaries, simply because none of them had ever done anything similar to what the test was all about (such as disassembling and cleaning an M-16)! Because none of them will ever do it, people from this secretarial group cannot be used as your group of non-masters.

2. 3/2 x 40 = 60 people altogether. Half should be masters (30), and half should be non-masters (30).
Do NOT let the number of available masters and non-masters in the tryout population dictate the number of items on your test. You MUST get enough people to test out the number of items you feel are necessary.

3.

	Non-Masters	Masters
Pass	6	26
Fail	24	4

Note that 10 people (16.7%) were incorrectly classified. Yes, this item seems to discriminate between masters and non-masters fairly well.

4.

	Non-Masters	Masters
Pass	13	18
Fail	17	12

There is a 50-50 chance of getting this item correct just by guessing, so you'd expect about 15 people out of 30 to get it right, by chance alone. And indeed, 18 of the masters got it right, and 13 of the non-masters got it right. Since only 3 more masters got it right than would be expected by chance, the item must be so difficult that it should be discarded.

5. Since so many non-masters got this item correct, the item should be omitted. It just didn't separate the masters from the non-masters.

Chapter 6

1. 1-i. 2-a. 3-h. 4-e. 5-j. 6-d. 7-l. 8-g. 9-c. 10-k. 11-b. 12-f.

13. Yes. Although the product was actually doing good repair work (so that the engine would indeed run smoothly), the process by which he achieved that product should also be noted by the examiner. And part of the process includes the trainee's careless behavior.
It's possible that the student could use some remedial practice in how he does repair work, even though he is able to perform the actual tuning and repairs successfully.

14. c. The trainee must pass the minimal number of items for each objective. You can't just add up the total number of items passed across all objectives, and then see if that value exceeds the criterion value for the overall test. Rather, each objective must be passed at some minimal level in order for the whole test to be passed.

Chapter 7

1. b. Think of reliability as the repeatability of test scores. Choices a and c refer to validity--does the test measure what it is supposed to measure? Choice d may help to increase reliability, but is not the correct answer here because it could refer to other things besides reliability.

2. a. If the test is really measuring what it's supposed to measure, then you should get about the same results when conducting a test-retest reliability check. Of course, external conditions and personal variables could decrease the reliability of the test results, as could confusion among judges about scoring procedures.

3. True. To take an oversimplified example, suppose that you thought that a baseball player's batting ability could be measured by (or predicted by, or was related to) his throwing ability. Certainly the maximum distance that he can throw a baseball will be a rather reliable measure over many such throwing trials. But the distance that he can throw a ball is not a valid measure (may not be highly correlated with) of his batting ability.

4. c. Validity will be increased because the test is a closer approximation to the "real thing." And higher fidelity means that irrelevant factors which might otherwise influence the performance of the test taker are reduced. Therefore, repeated performances should be more consistent. And the more consistent the performance, the higher the reliability.

5. People forget things over a period of time. And, some things that people learn since taking a test may interfere with the knowledge or skill that had been previously learned to pass the test.

6. $$\text{phi} = \frac{1 \times 17 - 10 \times 2}{\sqrt{19 \times 11 \times 3 \times 27}} = \frac{-3}{\sqrt{209 \times 81}} = -.02$$

Either conditions or personal variables (or both) were undesirable on the second day. Actually, the trainees were probably just too tired and poorly motivated to be taking a test at 1900.

7. $$\text{phi} = \frac{20 \times 5 - 1 \times 2}{\sqrt{21 \times 7 \times 22 \times 6}} = \frac{100 - 2}{\sqrt{19{,}404}} = +.70$$

There seems to be rather high concurrent validity.

8. $$\text{phi} = \frac{22 \times 2 - 5 \times 11}{\sqrt{27 \times 22 \times 33 \times 7}} = -.03$$

$$\text{phi} = \frac{3 \times 35 - 1 \times 11}{\sqrt{36 \times 4 \times 36 \times 4}} = +.72$$

$$\text{phi} = \frac{28 \times 2 - 6 \times 5}{\sqrt{34 \times 7 \times 33 \times 8}} = +.10$$

The first value of phi, -.03, is so low that there is very poor reliability for Form A test-retest reliability.
Examining the second and third phi coefficients, we may note that the Form A results from Monday correlate very highly with the Form B results from Friday. However, the Form A results from Wed. correlate very poorly with Form B results from Fri. What is the tester able to infer from all of this?
Well, something was probably quite unfavorable when Form A was given on Wed. Perhaps conditions or personal variables were adverse.
It therefore seems that Form A is reliable, Form B is also reliable, and that we can dismiss the results of Wed. as arising from adverse conditions external to the test.